THE PHYSICAL SOCIETY

Reports on
PROGRESS IN PHYSICS

VOLUME V

GENERAL EDITOR
ALLAN FERGUSON

Price to non-Fellows 20*s. net*

Reprinted by
WM. DAWSON & SONS LTD.,
with the permission of the Publisher

PREFACE

In those far-off days, more than thirty years behind us, when tariff reform was a live political issue, the pamphleteers of the day put forth a delightful volume describing Clara's adventures in the fiscal blunderland. The clock in Looking-glass World bore the features of Mr Arthur James Balfour, and as Clara looked at it, the clock struck, saying, "No set-tled con-vic-tions."

A carping critic might say that the phrase would serve to illustrate my own attitude to the scope and content of these Progress Reports. I prefer to think that I have preserved a certain elasticity of outlook and openness to argument, not attempting to fit the matter of the Reports into an over-rigid scheme. There have resulted, in consequence, slight variations from year to year in the content and outlook of the Reports which have not been altogether to their disadvantage.

The web of the volumes is "of a mingled yarn" but right through it has run one thread of policy—the endeavour to present articles which shall discuss the latest developments in physical science, and which shall at once make clear to the non-specialist the meaning and extent of these advances, and provide the researcher in the particular field under discussion with a technical résumé helpful to him in his own work. This has, on more than one occasion, been dismissed as an impossible task, and my grateful thanks are due to those authors who have so clearly and convincingly shown its practicability and usefulness.

Mr Awbery, my successor in the duties of Papers Secretary to the Society, now takes over the editorial work, and under his care the Reports are likely to enter on a career in which their usefulness to students of physical science will be still more clearly demonstrated.

ALLAN FERGUSON

Reprinted in The Netherlands

CONTENTS

PAGE

General physics 1

I. Surface tension. By R. C. BROWN

II. Brownian motion. By R. C. BROWN

III. Elasticity. By H. L. COX

IV. The viscosity of non-Newtonian fluids. By C. F. GOODEVE

The adsorption theorem of J. Willard Gibbs. Anomalies in applying it to surface tension curves that exhibit minima in dilute solution. By J. W. McBAIN and G. F. MILLS 30

Recent theories of the liquid state. By N. F. MOTT and R. W. GURNEY . 46

Plastics in industrial physics. By H. W. MELVILLE 64

Recent progress in astronomy. By A. C. D. CROMMELIN 82

Meteorology. By D. BRUNT 100

Sound. By E. G. RICHARDSON 121

Instrumental aids for defective hearing. By PHYLLIS M. TOOKEY KERRIDGE 150

Heat. By R. W. POWELL 164

Optics. *Editor*, L. C. MARTIN 182

I. General introduction. By the EDITOR

II. Optical instruments and the theory of optics. By L. C. MARTIN

III. Recent developments in optical glass manufacture. By W. M. HAMPTON and W. N. WHEAT

IV. Advances in the theory of photography. By E. W. H. SELWYN

V. Television optics. By W. D. WRIGHT

Spectroscopy. *Editor*, W. JEVONS 210

I. Atomic spectra. By Professor A. G. SHENSTONE

II. Hyperfine structure. By M. F. CRAWFORD

III. The Stark effect and some related phenomena. By J. STUART FOSTER

IV. Astronomical spectroscopy. By A. HUNTER

V. The spectra of the alkali-metal hydrides. By R. W. B. PEARSE

The soft x-ray spectroscopy of the solid state. By H. W. B. SKINNER . 257

The use of x-rays and γ-rays in medicine. By W. V. MAYNEORD . . 284

PAGE

Absolute electrical measurements. By L. HARTSHORN 302

Electric wave filters. By N. F. ASTBURY 334

Atomic physics 348

I. The single scattering of elementary particles by matter. By F. C. CHAMPION

II. Cosmic rays. By W. HEITLER

The mechanism of the Geiger counter. By A. NUNN MAY 390

Some difficulties and developments in quantum mechanics. By H. T. FLINT 402

The teaching of physics in schools. By A. W. BARTON 422

GENERAL PHYSICS

I. SURFACE TENSION

By R. C. BROWN, PH.D.

§ 1. THE SURFACE TENSION OF SOLUTIONS

In this section of the *Progress Reports* for last year mention was made of some unexpected results due to Jones and Ray[1] for the surface tension of aqueous solutions of inorganic salts. Until the publication of their paper it was universally supposed that dissolved inorganic salts raised the surface tension of water continuously as their concentration was increased from zero to saturation. Jones and Ray found, however, that for very small concentrations these substances actually lowered the surface tension of water, and the surface tension-concentration curve showed a definite minimum. Between 1924 and 1937 several theories of the surface tension of electrolytes were advanced, and references to them may be found in last year's *Reports*; but none of them had forecast the results of Jones and Ray. To their number we may now add the theory of Belton[2].

Dole[3] has, however, suggested a theory to account for what he calls the "Jones effect". He postulates the presence in the surface of "active spots" to which negative solute ions may be attracted. The number of such spots is small and they may consist of groups of water molecules whose proton ends are pointing inwards. Such a condition has previously been suggested by Frumkin[4] and Ariyama[5]. Clearly these regions would become areas of positive adsorption of (negative) ions, and for small bulk concentrations it might well be that this effect more than balanced the negative adsorption occurring all over the remaining surface area. Thus the net positive adsorption required by the Gibbs equation to account for the initial negative slope of the surface tension-concentration curve might be established. Following the method of Langmuir, Dole arrives at a relation between surface tension and concentration which fits the curves of Jones and Ray quite well, and, moreover, reduces to Szyszkowski's equation for capillary-active solutes when the whole surface is supposed capable of attracting solute molecules.

The problem of the change of the surface tension of solutions with time, mentioned previously in these *Reports*, has received some attention. Bond and Puls[6] have worked out a theory of this change purely from the point of view of the time taken for molecules of dissolved substance to diffuse from the interior to the surface. They predict that the time (τ) required for the surface tension to pass half way to its final value is given by

$$\tau = -\left(\frac{d\bar{\gamma}}{d\bar{c}}\right)^2_{\bar{c}\to 0} \Big/ DR^2T^2,$$

where $\bar{\gamma}$ is the static surface tension for a bulk concentration $\bar{c}$, D is the coefficient of diffusion, R the gas constant and T the absolute temperature. The basis of the theory is the Gibbs adsorption equation.

The authors then proceed to establish an equation which will give the value of τ from observations of the tension of a surface of any particular age. Measurements were made by the liquid disc method devised by Bond[7] and values of τ were calculated, the age of the surface at the time of its measurement having been estimated.

For pure liquids and solutions of substances such as ammonia, sodium chloride and potassium carbonate, Bond and Puls estimate that τ is of the order of 10^{-9} sec. and their experiments show that it is less than 0·005 sec. Solutions of "Lux" in water gave a predicted value of 0·18 sec. and observed values between 0·028 and 0·14 sec. according to concentration. Solutions of the alcohols gave good agreement between observed and predicted values, but in cases where τ is very large (e.g. decylic acid) there was considerable discrepancy. This may be due to association of molecules.

Adam and Shute[8] have published an account of surface-tension measurements made some years ago on aqueous solutions of paraffin chain salts. At low concentrations these substances show a very slow decrease of surface tension with time, and equilibrium is scarcely reached after several days. At higher concentrations, depending on the particular substance used, the rate of attainment of the steady state is greatly increased, and the change with time becomes scarcely detectable. If small quantities of inorganic salts (e.g. sodium bromide or sodium carbonate) are added to the solutions, the rate of decrease of surface tension is very greatly increased, the total time taken to reach the steady value being now only a second or so. The authors used the sessile bubble and maximum bubble pressure methods of measurement, and obtained similar but not completely comparable results with each. The writer of this article[9] also found similar effects with dilute solutions of the same substances using the ripple method.

Adam and Shute are of the opinion that a complete explanation of the phenomenon has not yet been found, but they emphasize certain points. First, it is noticed that the concentration at which the rate of fall becomes very rapid is that at which conductivity measurements indicate the formation of ionic micelles. Secondly, the slow change of surface tension is not due to factors which have been found to operate in other cases of similar behaviour, namely, the presence of carbon dioxide and preferential evaporation.

§ 2. CONTACT ANGLE, WETTING AND SPREADING

The proceedings of a conference on "Wetting and Detergency" have been published[10] and its pages clearly emphasize the importance of surface phenomena in industry. It is proposed, however, to make no more than this passing appreciative reference to the symposium, and to call attention to some papers which have appeared in the less technological journals.

Kneen and Benton[11] have made some interesting experiments on contact angles using a very direct method of measurement. An enlarged shadow of a liquid drop resting on a given plane surface was formed and its outline drawn. Drops of diameter not exceeding 0·3 cm. gave satisfactory results. Values obtained for water on paraffin wax and other solids gave good agreement with other authors. In order to evaluate the wetting capacity of soaps, measurements were made of the contact angles between paraffin wax and aqueous solutions of the sodium salts of a series of fatty acids. For salts containing up to 8 carbon atoms there was a progressive slight decrease of contact angle, and from 8 to 14 carbon atoms the decrease became very rapid. Extrapolation suggests that for 16 carbon atoms (sodium palmitate) the angle would be zero, i.e. there would be complete wetting. Solutions of 0·01 molar strength were used throughout.

Further, Kneen and Benton have made experiments with water on solid polar surfaces of stearic and palmitic acids. First of all non-polar surfaces of the acids were prepared by allowing them to solidify in air, and with these water gave a contact angle of about 105°. Angles of 60° or 70° were obtained however when polar surfaces were prepared by allowing solidification on glass and using the smooth surface which had been in contact with the glass. The authors have also concerned themselves with such obviously important and interesting matters as the wetting powers of insecticidal and fungicidal solutions on leaves and of various solutions on human teeth.

Bangham and Razouk[12] point out that when considering the equilibrium of a solid and liquid the effect of the saturated vapour should be taken into account. Thus for the equation of equilibrium they write

$$\gamma_{SV} = \gamma_{SL} + \gamma_{LV} \cos\theta,$$

where the subscripts S, L and V refer respectively to solid, liquid and vapour.

Under these conditions the work necessary to separate the solid from the liquid will be

$$\begin{aligned} W &= \gamma_{SV} + \gamma_{LV} - \gamma_{SL} \\ &= \gamma_{LV}(1 + \cos\theta). \end{aligned}$$

It is pointed out that W is not the "work of adhesion" as ordinarily understood, since the effect of the vapour is not usually considered.

The most important action of the vapour is to cover the surface of the solid which is exposed to it with an adsorbed layer, and Bangham and Razouk contend that this layer is not in the same state as the bulk of the liquid, since unlimited condensation does not take place on to it and there is often a finite contact angle within the bulk of the liquid which fails to spread even when the adsorbed film is thick enough to show interference colours[13].

In the same paper[12] the authors have established an expression for the total heat of wetting (or "diminution of heat content which accompanies immersion in the liquid") in terms of two distinct parts which represent respectively the heat evolved by the formation from the liquid of an adsorbed film in equilibrium with the saturated vapour, and the heat generated when such a surface is plunged into the liquid.

The adsorbed phase of common liquids on wood charcoal has been studied by the same authors[14].

Bangham and Saweris[15] have made an exhaustive study of the spreading of liquids on mica. They re-emphasize the importance of the adsorbed film formed from the vapour on the solid-vapour surface. Thus the energy of this surface will be different from that which it would possess *in vacuo*, its "effective" energy γ_{SV} being equal to $\gamma_{SO}-F$, where F is the "surface back pressure" due to the adsorbed film. Consequently the presence of saturated vapour often prevents spreading by assisting the formation of this layer. From the fact that the film may be built up so as to show interference colours the authors conclude that its thickness must amount to several thousands of Ångström units even while it retains a different character from the bulk liquid. Therefore the range of action of the mica surface is at least equal to this.

Doss and Rao[16] have written on somewhat similar lines. They show by means of a very simple theory that if the adsorbed layer on the solid-vapour surface covers only σ sq. cm. per sq. cm. of available solid surface, then

$$\cos\theta = 2\sigma - 1.$$

Thus the film is complete (i.e. $\sigma = 1$) only if $\theta = 0$. It would appear to the present writer that this statement is, in a sense, equivalent to the contention of Bangham and Saweris[15] that the condition of the film is not the same as that of the liquid in bulk, and the amount by which σ differs from unity is a measure of the degree of difference of the two states.

By taking the roughness of the solid surface into account, Doss and Rao show by means of their simple theory that, if $\theta < 90°$, increasing roughness will decrease θ, and if $\theta > 90°$ the reverse effect will be observed. Thus enhanced wetting of newly fused glass by water is brought about by washing which increases roughness by dissolving soluble matter, and mercury wets quartz more readily than it wets glass owing to the greater smoothness of the quartz.

A discussion of the role played by "work of adhesion" in the flotation process and the measurement of its value for certain systems will be found in a paper by Valentiner[17].

Burdon, Fuller and Gibson[18] have followed the spreading of water on mercury by means of a cinematograph camera. Small quantities of acid were found to accelerate spreading and to cause the drop to reach a final area proportional to the acid concentration. Conductivity water was found to spread slowly on mercury, its diameter increasing linearly with time. Alkalis in water inhibit spreading and it was found that they increased the interfacial tension with mercury, whereas acids decreased it.

Allen, Grant and Burdon[19] have found that pure heavy paraffins spread on water if they are irradiated with ultra-violet light either after being placed on the surface or some months before the experiment. It was found that, after the diameter of the drop had reached about 1 cm., the edge moved outwards with uniform acceleration.

Ramdas and Vaidyanathan[20] have carried out some interesting experiments on the spreading of camphor on water. They show that a piece of camphor floating on water loses weight chiefly by *surface* solution. In their experiments the rate of loss of weight was greatly diminished if the surface was covered with an oil or if the camphor was completely submerged. Further experiments showed that the rate of surface solution was proportional to the length of the line of contact between camphor and water and independent of the area of contact. The available area of water surface apparently had no influence on the rate of solution, and it was shown that when the area was small the film of camphor which was continually spreading from the floating solid turned downwards when the edge of the vessel was reached, and set up quite observable convection currents. Ramdas and Vaidyanathan consider that these convection currents are responsible for another fact which they found, namely, that although water carrying a soluble surface film evaporates *more* rapidly than when it has a clean surface, yet no temperature difference between the surface and the interior was noticeable in the former case, whereas in the latter it was, of course, considerable.

§ 3. DROPS

Riedel[21] calls attention to the behaviour of a drop of sodium oleate solution falling on to a weak soap solution. He favours the theory which postulates an air sheath surrounding the drop, and describes how, for a certain height of fall, the drop may be observed to enter the liquid and remain below the surface without mixing. By means of a simple funnel arrangement he was able to collect and measure the air which surrounded the drop when it was thus suspended. The thickness of the air film turns out to be of the order of 10^{-4} cm.

A more critical study of liquid drops on liquid surfaces has been carried out by Hazlehurst and Neville[22]. These authors have measured the "critical height", i.e. the maximum height from which a drop may fall in order to remain stable on the surface of the second liquid, and they find that the more nearly alike the drop and bulk liquids are the less likely are they to merge instantly. This was shown particularly when benzene was dropped on to mixtures of benzene and ethanol, no stable drops being formed if the concentration of benzene fell below 90 per cent. In nearly all cases a liquid formed drops on itself more readily than on another.

Hazlehurst and Neville consider that the hypothesis of an air or vapour film separating the drop from the liquid upon which it rests is untenable for several reasons. For example, a flow of liquid from the drop to the bulk can sometimes be observed. The theory of a repulsive electrical force also seems unsatisfactory, and the authors suggest that, in the case of polar liquids, the molecules in the surface of both the drop and the supporting liquid are oriented with their weakest fields outwards. These fields then fit each other and there is very little residual force. Thus the greatest stability is obtained when the molecules in the bulk and the drop are identical, or nearly so, and, further, the most stable drops are observed, as might be expected, when the liquids consist of long polar molecules. The chance of surface molecules becoming reversed and consequently being attracted across the

barrier is not so great in the case of long molecules which are strongly oriented and tightly packed by forces tangential to the surface.

The study of falling drops by means of high speed cinematography has been continued by Edgerton, Hauser and Tucker[23]. In the first place these authors have investigated the relation between the mechanism of drop formation and the shape of the empirical correction curve due to Harkins and Brown [24], which is a graph relating the drop weight expressed as a fraction of the "ideal drop" (a curious conception whose weight is defined as $2\pi r\gamma$) with the quotient of the radius of the tip, r, by the cube root of the volume of the drop. The curve shows both a minimum and a maximum value of the fractional drop weight; it appears that the decrease to the minimum as r increases is due to the decrease of the radius of the drop relative to r, and the subsequent rise to the maximum may be due to the increasing relative size of the stem which joins the drop to the liquid remaining on the jet at the instant of rupture. It is this stem which gives rise to the secondary drops. The authors' photographs show progressive changes in the mode of detachment as the diameter of the tip increases from 0·2 to 1·6 cm.

Sodium oleate in water shows two properties in common with other long-chain compounds. Its surface tension requires a considerable time to fall to an equilibrium value, particularly in the case of dilute solutions, and the curve of static surface tension against concentration shows a minimum. It is found that this minimum corresponds to a marked change in the method of detachment of the drop, and, since complete maturity of the surface cannot be reached during the dynamic process of drop detachment, it follows that the transition exists even when the surface is not completely aged.

Lastly Edgerton, Hauser and Tucker show that if a number of solutions are made up so as to possess equal surface tensions (as shown by the du Noüy method) but to have widely different viscosities and densities, then the mechanism of drop formation is also very different. The authors therefore suggest that the use of the drop-weight method with any but pure non-viscous liquids is open to objection.

§ 4. MISCELLANEOUS RESEARCHES

Prof. Porter's work on the calculation of surface tension from experiment has been mentioned from time to time in these *Reports*, and this year we note that his latest paper[25] supplies data which make it possible to calculate the value of the surface tension of a sessile drop of liquid from observations of its dimensions whatever the size of the drop may be.

A relation between the solubility of gases in a liquid and the surface tension of the liquid has been suggested by Uhlig[26]. If a spherical gas molecule of radius r enters the solvent it will produce a cavity, and consequently $4\pi r^2$ cm.^{2} of liquid-gas interface with an energy of $4\pi r^2\gamma$ ergs will be created. Other energy changes may take place on solution owing to interaction between solvent and solute molecules, and if these are denoted by E, then the total energy change, Δu, is given by

$$\Delta u = 4\pi r^2\gamma - E.$$

If η is the Ostwald coefficient of solubility, i.e. the ratio of the concentration of gas in the solvent to the concentration of gas in equilibrium with it, then

$$\eta = \frac{c'}{c} = e^{-\frac{\Delta u}{kT}},$$

using the Maxwell-Boltzmann distribution theorem.

Therefore $$\log \eta = -\frac{4\pi r^2 \gamma - E}{kT}.$$

Thus we should expect a straight-line relationship between γ, the surface tension of various liquids, and η, the solubility of a particular gas in those liquids, provided the conditions are such that E and its variation from one liquid to another are small. Experimental results for CH_4, O_2, N_2 and H_2 in ethyl ether, acetone, methyl acetate, carbon tetrachloride, benzene and chlorobenzene confirm the linear relation. For any particular gas the slope of the graph gives its molecular radius in good agreement with values obtained from van der Waals's constants. E may be calculated from the intercept and is found to be negative if the solubility is low and positive if it is high.

A certain amount of attention has lately been directed towards the significance of Macleod's equation,

$$\gamma = C(D-d)^4,$$

where C is a constant practically independent of temperature, and D and d are liquid and vapour densities respectively. The parachor, P, invented by Prof. Sugden is defined by

$$P = MC^{\frac{1}{4}},$$

where M stands for molecular weight. Using an equation due to Katayama[27] and the law of rectilinear diameters, D. T. Lewis[28] has arrived at the following expression for C:

$$C = k(T_c - T)/M^4 [(4T_c - 2T)/V_c T_c]^{\frac{1}{2}}.$$

The author points out however that this equation must be erroneous, since it makes C become zero at the critical temperature which is contrary to observation, and the following relation is suggested:

$$\gamma = [(T_c - \alpha)/MK]^4 (D-d)^4,$$

where "the constants α and K remain the same for one distinct chemical group and are probably configurational or cohesional functions characterizing a particular type of molecule."

We then have $$P = (T_c - T)/K,$$

i.e. we should expect a straight-line relationship between parachor and critical temperature for any particular group of substances. This is found to be the case.

A graphical test also shows that, at any rate for the halogens (including halogen halides) and the inert gases, there is also a linear relationship between parachor and *boiling point*.

It is interesting to compare the above analysis with one of similar scope but different viewpoint published by Ferguson and Kennedy[29] and noticed in Vol. IV of these *Reports*.

Some interesting experimental results have been recorded by these last-named authors[30] who have now measured the surface tension of liquid crystals at various temperatures near their transition points. The method used was that suitable for small quantities of substance and previously described by them[31], and the substances measured were anisaldazine, *p*-azoxyanisol and *p*-azoxyphenetol.

The first of these had previously been measured by F. M. Jaeger[32] using the maximum bubble pressure method, but, by taking small temperature intervals, Ferguson and Kennedy have been able to demonstrate the interesting changes of surface tension which occur at the transition point. All their curves show that in the neighbourhood of this temperature the surface tension *rises* during an increase of 4 or 5° C. and then begins to fall again at a rate different from that prevailing below the transition temperature. Thus their curves assume what they call a "van der Waals" shape.

The use of surface-tension measurements in the identification of proteins present in certain biological fluids is described by Perryman and Selous[33] in a paper which is primarily of biochemical interest.

REFERENCES

(1) Jones, G. and Ray, W. A. *J. Amer. Chem. Soc.* **59**, 187 (1937).
(2) Belton, J. W. *Trans. Faraday Soc.* **33**, 1449 (1937).
(3) Dole, M. *J. Amer. Chem. Soc.* **60**, 904 (1938).
(4) Frumkin, A. *Z. phys. Chem.* **111**, 190 (1924).
(5) Ariyama, K. *Bull. Chem. Soc. Japan*, **12**, 109 (1937).
(6) Bond, W. N. and Puls, H. O. *Phil. Mag.* **24**, 864 (1937).
(7) Bond, W. N. *Proc. Phys. Soc.* **47**, 549 (1935).
(8) Adam, N. K. and Shute, H. L. *Trans. Faraday Soc.* **34**, 758 (1938).
(9) Brown, R. C. *Proc. Phys. Soc.* **48**, 312 (1936).
(10) *Wetting and Detergency*. Published by A. Harvey, London (1937).
(11) Kneen, E. and Benton, W. W. *J. Phys. Chem.* **41**, 1195 (1937).
(12) Bangham, D. H. and Razouk, R. I. *Trans. Faraday Soc.* **33**, 1459 (1937).
(13) Bangham, D. H., Mosallam, S. and Saweris, Z. *Nature, Lond.*, **140**, 237 (1937).
(14) Bangham, D. H. and Razouk, R. I. *Trans. Faraday Soc.* **33**, 1463 (1937).
(15) Bangham, D. H. and Saweris, Z. *Trans. Faraday Soc.* **34**, 554 (1938).
(16) Doss, K. S. G. and Rao, B. S. *Proc. Indian Acad. Sci., Bangalore*, A, **7**, 113 (1938).
(17) Valentiner, S. *Kolloidzschr.* **83**, 1 (1938).
(18) Burdon, R. S., Fuller, G. R. and Gibson, E. S. H. *Trans. Faraday Soc.* **33**, 1528 (1937).
(19) Allen, W. D., Grant, Kerr and Burdon, R. S. *Trans. Faraday Soc.* **33**, 1531 (1937).
(20) Ramdas, L. A. and Vaidyanathan, P. S. *Proc. Indian Acad. Sci., Bangalore*, A, **7**, 186 (1938).
(21) Riedel, W. *Kolloidzschr.* **83**, 31 (1938).
(22) Hazlehurst, T. H. and Neville, H. A. *J. Phys. Chem.* **41**, 1205 (1937).
(23) Edgerton, H. E., Hauser, E. A. and Tucker, W. B. *J. Phys. Chem.* **41**, 1017 (1937).
(24) Harkins, W. D. and Brown, F. E. *J. Amer. Chem. Soc.* **41**, 499 (1919).
(25) Porter, A. W. *Phil. Mag.* **25**, 752 (1938).

(26) Uhlig, H. H. *J. Phys. Chem.* **41**, 1215 (1937).
(27) Katayama, M. *Sci. Rep. Tôhoku Univ.* **4**, 373 (1916).
(28) Lewis, D. T. *J. Chem. Soc.* p. 261 (1938).
(29) Ferguson, A. and Kennedy, S. J. *Trans. Faraday Soc.* **32**, 1474 (1936).
(30) Ferguson, A. and Kennedy, S. J. *Phil. Mag.* **26**, 41 (1938).
(31) Ferguson, A. and Kennedy, S. J. *Proc. Phys. Soc.* **44**, 511 (1932).
(32) Jaeger, F. M. *Z. anorg. Chem.* **101**, 1 (1917).
(33) Perryman, P. W. and Selous, C. F. *J. Physiol.* **92**, 151 (1938).

II. BROWNIAN MOTION

By R. C. BROWN, Ph.D.

In this section, which reviews the work of the past three or four years, it is proposed to mention chiefly those papers which have dealt with the experimental aspects of Brownian motion. A number of mathematical works on the kinetic theory of gases have included references to Brownian motion, and their content can quite well be learnt from the abstracting periodicals.

A new light source devised by Hoffmann[1] for ultramicroscopy will be of interest to experimenters. It consists of a much over-loaded Osram mercury lamp, and it enables particles in motion to be photographed with exposures of the order of one-tenth of a second.

The question of the observation of Brownian movement by the naked eye has received some attention. Following papers by Andrejev[2] and Lau and Johannesson[3], Obrist[4] has pointed out that the first observations of this kind were made by Molisch[5] who used the chyle or milk of a plant (*Euphorbia splendens*). This liquid, when mounted between a microscope slide and coverslip and viewed obliquely in transmitted sunlight, exhibits continually changing interference colours. Kappler[6] has made a more thorough study of this phenomenon from the theoretical point of view. He summarizes the contrasting conditions which are necessary for visibility with and without the ultramicroscope. For instance, in the latter case the suspension should have a fairly high concentration such as might cause the particle separation to be unresolvable in the microscope. Kappler considers that the effects observed by the unaided eye are due to diffraction by fortuitous collections of particles which act as irregular gratings. The changes with time are due to the continuous re-arrangement of the particles by Brownian movements. The average distance between the molecules must be less than the limit of resolution of the eye.

The action of light on Brownian movement has also been the subject of some discussion. Pospíšil[7,8,9] considers that the effect of polarized light is to cause an enhanced movement in the direction of the magnetic vector as compared with that parallel to the electric vector. Fürth and Zimmermann[10] have made careful experiments both with unpolarized and polarized light. In the former case no modification of the movement was observable, and in the latter case the mean square

displacement in the magnetic direction was 7–11 per cent greater than that in the electric direction which was found to equal that in the direction of propagation of the light. Filtering out the short waves diminished the effect. Fürth[10] considers that what actually occurs is a *diminution* of the movement both in the direction of propagation and in that of the electric vector whilst the movement in the magnetic direction is unchanged. He favours neither the theory of Pospíšil[11], who postulates a transference of light energy to the particles by collisions of the second kind, nor that of Deubner[12] who appeals to photophoresis for an explanation. Fürth himself suggests an explanation of the effects as described by himself and Zimmermann on the basis of photoelectric action. The photoelectrons liberated from the surface of the particle form an ionic cloud which decreases the mobility of the particle in the direction of propagation of the light and also parallel to the electric vector, but not in the magnetic direction.

Kolb[13] has made experiments which seem to confirm Fürth's conclusions. Using a gold sol, the particle size of which he measured by two independent methods (direct counting and the "fall" method), he found that in the direction of the light and of the electric vector the movements were equal, and they were both 20 per cent less than that in the magnetic direction. The calculated movement for normal conditions was found to agree with that in the *magnetic* direction under the influence of the polarized light. In the notes which follow Kolb's paper Deubner[14] draws attention to possible objections to Kolb's experiments and to Fürth's theory, which, he says, does not explain the lack of action of unpolarized light on the Brownian motion. Fürth[15] dismisses the criticisms of Kolb's experiments and considers that his (Fürth's) theory is as satisfactory as any.

The action of a magnetic field upon Brownian motion is a question which is by no means settled. Svedberg[16], in 1912, considered that there was no action, and recently Métadier[17] has failed to discover any with fields up to 20,000 gauss, although he used in turn ferro-, para-, and non-magnetic particles. On the other hand, Schaefer[18] has shown in a purely theoretical paper that a modification of Brownian motion should be observable on account of the heating which an electrically conducting particle would experience if it were in motion in a magnetic field. The gain of heat would take place at the expense of the kinetic energy of the particle.

Turning to the Brownian motion of suspended systems we find that Niessen[19] has worked out a theory of the motion of a critically damped galvanometer. He indicates how experiments with such an instrument would lead to a value of Boltzmann's constant.

Kappler[20] has continued his work on the Brownian motion of a torsion balance and has produced some interesting results. His balance had a moment of inertia of the order of 10^{-7} g.-cm.^{2} and a time period of about 15 sec. It was surrounded by gas at a pressure of a few hundredths of a mm. of mercury, and its movements were recorded by means of a spot of light on a continuously moving film. Analysis of the traces gave the displacement and velocity of the suspended system at any instant and enabled the distribution of these quantities to be examined.

The Maxwell-Boltzmann distribution law was verified and a satisfactory value for Boltzmann's constant was calculated from the equation

$$\tfrac{1}{2}m\overline{u^2} = \tfrac{1}{2}kT,$$

where m is the moment of inertia of the suspended system and $\overline{u^2}$ the mean square angular velocity. In a second paper Kappler[21] discusses a theory of errors applicable to his experiment.

Gerlach[22] has published an interesting article on the theoretical limits of measurability. He explains clearly how, if a very sensitive balance is subject to Brownian motion, the error in the estimation of a small load by means of the observation of successive swings is reduced by increasing the time during which continuous observations are made.

Perrin[23] has worked out the Brownian movement of rotation of an ellipsoidal particle both when free and when under the action of an alternating field. His results enable an extension to be made to the Debye theory of dielectric dispersion so as to include the case of non-spherical molecules. In a further paper Perrin[24] puts forward a theory of the polarization of the fluorescence emitted by solutions containing non-spherical molecules.

REFERENCES

(1) Hoffmann, K. *Kolloidzschr.* **83**, 9 (1938).
(2) Andrejev, N. *C.R. Acad. Sci., U.R.S.S.*, **1**, 62 (1934).
(3) Lau, E. and Johannesson, J. *Phys. Z.* **35**, 1013 (1934).
(4) Obrist, J. *Phys. Z.* **36**, 343 (1935).
(5) Molisch, H. *S.B. Acad. Wiss. Wien*, Math.-Nat. Kl. **116** (Abt. 1), 467 (1907).
(6) Kappler, E. *Phys. Z.* **36**, 643 (1935).
(7) Pospíšil, V. *Ann. Phys., Lpz.*, **83**, 735 (1927).
(8) Pospíšil, V. *Ann. Phys., Lpz.*, **23**, 259 (1935).
(9) Pospíšil, V. *Phys. Z.* **31**, 65 (1930).
(10) Fürth, R. and Zimmermann, O. *Ann. Phys., Lpz.*, **24**, 183 (1935).
(11) Pospíšil, V. *Phys. Z.* **31**, 78 (1930).
(12) Deubner, A. *Phys. Z.* **31**, 909 (1930).
(13) Kolb, J. *Phys. Z.* **39**, 194 (1938).
(14) Deubner, A. *Phys. Z.* **39**, 198 (1938).
(15) Fürth, R. *Phys. Z.* **39**, 198 (1938).
(16) Svedberg, The. *Die Existenz der Moleküle* (Leipzig, 1912).
(17) Métadier, J. *C.R. Acad. Sci., Paris*, **199**, 1196 (1934).
(18) Schaefer, C. *Ann. Phys., Lpz.*, **32**, 190 (1938).
(19) Niessen, K. F. *Physica*, **3**, 841 (1936).
(20) Kappler, E. *Ann. Phys., Lpz.*, **31**, 377 (1938).
(21) Kappler, E. *Ann. Phys., Lpz.*, **31**, 619 (1938).
(22) Gerlach, W. *Z. Ver. dtsch. Ing.* **81**, 2 (1937).
(23) Perrin, F. *J. Phys. Radium*, **5**, 497 (1934).
(24) Perrin, F. *J. Phys. Radium*, **7**, 1 (1936).

III. ELASTICITY

By H. L. COX

The theory of elasticity, as distinct from the theory of strength of materials, is entirely concerned with special solutions of certain rather complex differential equations. Progress is to be measured partly by the derivation of new solutions appropriate to problems of practical interest, but more particularly by the development of new methods by which analysis of complicated problems may be attempted. Such methods must in general be only approximate; but, if they permit solution by successive approximation, they are as valuable in practice as exact methods.

The method of relaxations which was briefly described in the *Reports on Progress in Physics*, 1936 (p. 17), affords a good instance where solution by successive approximation is actually preferable to solution by exact methods. To analyse by the exact method the load distribution in a highly redundant framework, it is first necessary to write down and solve a set of simultaneous equations equal in number to the number of redundancies. Theoretically no special difficulty is involved, but if the number of redundancies is large, the derivation of the equations is laborious and their solution to a sufficient degree of accuracy by ordinary methods may be virtually impossible.

Purely mathematical methods for the solution of a large number of simultaneous equations by successive approximation have of course been devised; but such methods have usually been developed for application to special classes of problem (e.g. Gauss's method in the theory of errors) and the coefficients have usually to conform to certain rules. Moreover, in order to apply a method of this kind, the simultaneous equations themselves must first be derived. The relaxation method in effect avoids both these steps by the expression in symbols of a possible *experimental* process of successive approximation, and in its first applications confidence in the convergence of the method was founded only upon the apparent convergence of the experimental procedure. A rigid mathematical proof that the method is essentially convergent has yet to be developed; but a proof based on considerations of energy has been stated by Black and Southwell[1].

Since the separate solutions obtained in succession at each relaxation of a constraint are finally added, it might appear that the method depends upon the principle of superposition; but, if the analytical process really represents the (possible) experimental procedure, it is clear that this restriction should be unnecessary. In fact, the method is applicable to structures involving end loaded members, to which the principle of superposition is not applicable, provided the effect of the end load in a member in modifying its resistance to side load is taken into account at each stage[45]. In such cases, however, the process is necessarily more laborious and, if the compressive end load approaches the Euler buckling valve, the convergence may become very slow.

For practical purposes simple convergence is scarcely sufficient; it is necessary also that the approach to the limit should be reasonably rapid. The first application of the method by Hardy Cross was to the discussion of end fixing moments in building structures of which the shear rigidity was due entirely to the rigidity of the joints. Now any member of uniform flexural rigidity, which shares in the rotation of a joint consequent upon the relaxation permitting rotation and which by that rotation transfers a couple M away from the joint, transmits it partly as a couple $-\frac{1}{2}M$ applied to its far end and partly as a couple $3M/2$ in the form of equal and opposite shearing forces at its two ends. Thus, if attention is restricted to moments, each clamping moment is halved at each relaxation and the process is rapidly convergent[46]. In consequence of this property the method is particularly useful in application to the problem of a beam continuous over several rigid supports[45]. On the other hand, forces transferred from one end of a member to the other suffer no diminution, so that the rapidity of the elimination of the forces on the (imaginary) constraints depends entirely on the transference of these forces to the actual points of support. It is in this respect that only slow convergence may be obtained.

If two or more joints none of which is actually supported are joined by members which are rigid in comparison with the other members joining these joints to the actual points of support, successive relaxations of the imaginary constraints at these joints will result primarily in the transfer of load to and fro along the more rigid members[45]. To avoid this disadvantage, various special devices have been suggested, all of which aim at initially reducing the total load to be so transferred. Southwell's device of "block relaxations" provides master constraints which include whole groups of such joints. By relaxation of the master constraint, leaving the individual joints of the group constrained to one another, rapid transference of the direct loads to the actual points of support can be effected. The preliminary adjustments described by Baker and Ockleston achieve the same purpose by very similar means.

The application of the relaxation method to beams of variable cross-section loaded in any arbitrary manner has been described by Bradfield and Southwell[2]; but Black and Southwell[1] have pointed out that the method is applicable to any physical problem where a similar experimental procedure can be conceived. They have shown how the method may be applied to any problem which involves the minimization of a homogeneous quadratic function of variable parameters and have illustrated this generalization by consideration of the adjustment of errors in surveying and of the distribution of current in a non-inductive network. Problems of this kind may be termed "energetic" in that the homogeneous quadratic expression often represents energy. Such problems are characterized by certain reciprocal relations between the coefficients of the simultaneous equations (e.g. Maxwell's reciprocal theorem in elasticity). Black and Southwell have shown[1] however that a modified relaxation method is applicable to other types of problem characterized by skew symmetry and have illustrated this extension by application to the current distribution in an inductive network; the convergence of the method in these cases is however not generally established.

Other contributions to the theory of frameworks have been made by Jagn[3] and by Broglio[4]. Jagn has drawn attention to an analogy (first noticed by Mohr) between the distribution of strain in a loaded structure and the distribution of load in a conjugate structure uniquely defined in relation to the actual structure; he has indicated how this analogy may be used to solve structural problems, but the full possibilities of the method have yet to be explored. Broglio has developed a general theory of plane frames and has described a new method based on the theory of matrices for calculating the distribution of load.

Any problem of elasticity is governed by four separate sets of conditions, the equations of equilibrium of stress components, the conditions for compatibility of strains, the relations between stress and strain and the conditions determining the boundary values of the stresses and/or displacements. The first set provides that each particle of the body shall be in static equilibrium, the second that the body shall remain a continuous medium, and the fourth that the body shall conform to the impressed configuration and/or be in equilibrium with the impressed forces. The third set differs in type from the others in that it depends on the actual properties of the material, but so long as the state of stress at a point is a unique function of the state of strain or vice versa, the exact nature of these relations is not of primary importance. The first three sets of conditions apply throughout the volume of the body, and one method of solution is to combine these three sets into one or more differential equations and to seek solutions of these equations which shall satisfy the boundary conditions. Only in specially simple cases, however, can solutions exactly fulfilling all the boundary conditions be found, so that in practice a solution found in this way represents the true solution of the stated problem with slightly different boundary conditions. The relaxation method is an example of this type of solution; the imaginary constraints impose artificial boundary conditions and the process of approximation consists in the gradual modification of these boundary conditions towards identity with those actually postulated.

In several important classes of elastic problems, notably the problems of torsion and flexure of prismatic bars and the problems of plane stress and plane strain, simple solutions appropriate to certain special boundary conditions can be found. In the flexure and torsion problems, for instance, such solutions satisfy the condition that the surface of the cylinder or prism should be free from tractions but require that the loads at the ends should be distributed in a special manner, whilst in the case of plane stress the solutions require that the edge loads should all be uniformly distributed over the thickness of the plate (of uniform thickness). The number of solutions of this type is continually being increased and reference can be made only to one or two of the solutions recently published. The solution of problems of this type depends on the determination of plane harmonic functions which shall assume assigned values over the perimeter or perimeters of the cross-section. Such solutions can be found in various ways; Shepherd[5] has used special functions to solve the problem of flexure for a beam having a cross-section in the form of a cardioid, whilst Seth[6] has developed a general solution for a prism of polygonal cross-section by the method of conformal transformation. This latter method is of

great assistance in problems of this kind, and Bradfield, Hooker and Southwell[7] have recently described how the method may be applied experimentally by the use of an electric tank. By effecting an improvement in the boundary conditions, they have shown how the method may be used to solve the torsion problem for a bar of triangular section. The late Prof. Filon[8] has pointed out a third solution analogous to St Venant's solutions for flexure and torsion; this solution represents the deformation of a cylinder stressed by axial shears over its surface balanced by uniform tension over its cross-section and represents the case of a long cylinder pulled through a close fitting hole. The torsion of curved beams, which finds an important application in helical springs, has been treated by Sen[9] and others; whilst Fish[10] has pointed out analogies between this problem and problems of lubrication and electric currents. Numerous additions to the solutions of problems of plane stress have also been made, the solution by Chitty and Pippard[11] of the stresses in a disc wheel is a typical example. Tolotti[12] has discussed the use of Airy's stress function (see *Report*, 1936) for multiple connected regions, whilst a generalization of the plane stress solutions by which the shear stresses in directions normal to the plate are no longer required to be zero through the whole thickness of the plate has been described by Southwell[13]. Several of the papers quoted illustrate the solution of these classes of problem by series, and here again particular interest is attached to methods of determining convergent series which approximate to the true solution. This problem has been treated in general terms by Soboleff[14], who has described a mathematical procedure for determining such series, and similar methods have been applied by Smirnov[15] and others to problems of wave propagation. A general solution of the plane stress problem for a rectangular boundary has been stated by Beschkine[16].

If solutions of the type discussed above were applicable only when the boundary conditions were exactly fulfilled, their practical value would be small; but their application is greatly extended by St Venant's principle. This principle states that the effect of a self-equilibrated load system applied to a small area of an elastic body is largely confined to the immediate neighbourhood of that area. As a corollary, using the principle of superposition, it may be deduced that the stresses and strains set up in an elastic body by two statically equipollent load systems differ appreciably only in the regions close to the applied loads. This principle, which was stated by St Venant without proof, has recently attracted considerable attention. A general proof has been stated by Zanaboni[17], who has also examined the magnitude of the differences due to different distributions of applied load, and Goodier[18] has estimated from considerations of minimum energy the volume of material within which the effects of a self-equilibrated stress system should be sensibly confined. Basing his argument also on minimum energy, Southwell[19] has shown that the St Venant solutions for flexure and torsion may be derived by a direct chain of reasoning in contrast to the somewhat tentative nature of the original proofs.

The principle of minimum strain energy, which thus provides a logical proof in a case where the ordinary method can appeal only to inherent probability, is a particular case of the principle of minimum potential energy, which is itself a

particular case of the principle of least action. The more general principle, particularly as Rayleigh's Principle, is widely used in the theory of vibrations, and both Rayleigh's principle and the principle of minimum potential energy are often applied to problems of stability; but the use of the principle in ordinary problems of elasticity has been somewhat neglected. Since the equations of equilibrium of stress components are particular statements of the laws of motion which are embodied in the principle of least action, it might be anticipated that these equations could be deduced from the principle of least action by assuming the conditions for compatibility of strain components and the existence but not necessarily the form of the strain energy function; this is in fact the case. Southwell[19] has now shown that, if the strain energy function is expressed in terms of stress components in such forms that the equations of equilibrium are identically satisfied and if Hooke's law is assumed to hold, application of the principle of minimum strain energy leads to the conditions for compatibility of strain components. It is interesting to compare this result with the proofs of uniqueness stated by the late Prof. Filon[20] and by Seth[21] of the solutions obtained from the differential equations of equilibrium expressed in terms of stress components only.

In general, if the strain energy, expressed in terms of the stresses or strains as entirely arbitrary functions, is rendered a minimum by the variational method, the differential equations of equilibrium will be derived; but in addition the boundary conditions necessary to complete the specification of the problem will also result. An interesting example of this use of the principle of minimum energy is afforded by a paper by Miles[22] on the stability of rectangular plates supported on elastic beams. The solution in this example is completed by solution of the differential equations in a manner similar to that used by Bateman[47] for discussing the bending of such plates by normal forces; this latter analysis has recently been extended by Costello[23]. On the other hand, the principle can often be used to greater advantage in the same manner as Rayleigh's principle, that is for the purpose of selecting from a number of possible configurations that which by rendering the strain energy (or the total potential energy) a minimum affords the closest available approach to the true configuration. The possible configurations must be so chosen that either the conditions for compatibility of strains or the stress equations of equilibrium are identically satisfied and preferably they should also satisfy identically some if not all of the boundary conditions. By variation of the assumed configuration in such a manner as to render the total potential energy a minimum, whilst at the same time fulfilling, preferably exactly or at any rate approximately, any boundary conditions not satisfied identically, the remaining condition of equilibrium of stresses or of compatibility of strains can be satisfied as closely as the assumed configuration permits. One method is to assume expressions for the displacements (thus satisfying identically the conditions for compatibility of strains) in the form of series of similar terms (polynomials, harmonic series, etc.), and, if the assumed form is capable of representing the true solution, this true solution will be obtained. On the other hand, if the assumed form is incapable of representing the true form exactly, the solution obtained will be only approximate, and the

differential equations of equilibrium will also be only approximately satisfied. The method must therefore have much in common with other more general methods of approximate solution of differential equations. For instance the method is closely related to Galerkin's method, the application of which to problems of elasticity has recently been described by Duncan[24], and it may also be compared with the least squares method discussed by Frazer, Jones and Skan[25]. Further developments in the application of the energy and related methods are to be expected; but the value of the method is already well illustrated by several of the large number of papers on vibration and stability published in the last two years. Reference can be made only to a few of these papers, chosen principally as illustrations of method rather than on account of the special interest of the results. Methods and results in the theory of vibrations have been reviewed by Klotter[26] and a similar review of the theory of stability has been made by Tagliacozzo[27]. Barta[28] has pointed out how the energy method may be used to determine both upper and lower limits to the frequency of vibration of a membrane; Barta's method may be compared with that described by Temple and Bickley (see *Report* for 1935). Special problems of the transverse vibration of beams have been treated by the energy method by Balinkin[29], whilst the relative accuracy of the energy and other methods in this application has been very fully explored by Davies[30]. The vibrations of rectangular plates with free edges has been considered by Pavlik[31], who has shown that the calculated frequencies agree reasonably well with experimental values, although the calculated values for the higher frequency modes tend to be too high. The energy method has been used by Procter[32] for the discussion of several practical types of instability, and by Voinovsky-Krieger[33] for consideration of the stability of an elliptical plate with clamped edges subjected to uniform compression in its own plane.

A comparison between the exact and approximate methods is afforded by a paper by Leggett[34] on the stability of a plane rectangular plate compressed in its own plane by two equal and opposite concentrated loads, and the use of the exact method is further illustrated by papers by Green[35] on the stability of a thin twisted strip and by Sezawa and Watanabe[36] on the buckling of a rectangular plate with clamped edges subjected to uniform compression in its own plane.

The theory of vibrations of solids is a particular branch of the general theory of solid bodies in accelerated motion. The solution of the general problem must consist of a particular integral appropriate to the impressed forces and a complementary function which enables the initial conditions to be fulfilled. This complementary function must be composed of an infinite series of terms representing the infinite number of possible modes of vibration of the continuous elastic medium. Now, in general the ratios of the frequencies of oscillation in these separate modes will be incommensurable and the complementary function will in consequence be not truly periodic. On the other hand, functions of this kind have many properties analogous to those of truly periodic functions and it has proved possible mathematically to define a class of such functions termed "almost periodic". The mathematical study of the properties of almost periodic functions was commenced only recently (1924) and their application to the dynamics of elastic bodies has not

yet been well explored. A brief description of the elementary properties of the functions was given by Titchmarsh[37] in 1929, but more recent work does not appear to have been summarized. An example of the application of the functions to a problem of elasticity is given in a paper by Avakian[38]. The problem of the elastic body in accelerated motion has also been considered by Pugno[39] and by Fairthorne[40].

Another important class of elastic problems is that of thin shells. Southwell[41] has recently shown how the problem of the torsion of a conical shell can be solved by considering the equilibrium under shearing forces of the shell developed into a plane trapezium. He has pointed out that, whereas in a parallel tube of other than circular section the cross-sections warp under torsion, in the tapered tube such warping would involve displacements of parts of the tube along the generators. He concludes that such warping cannot occur and that distortion of the cross-sections by displacements perpendicular to the generators must take its place. The torsion of thin walled tubes of rectangular section has been investigated both theoretically and experimentally by Hovgaard[42]. He has found that although the overall twist measured experimentally agrees very well with theoretical predictions, the actual distribution of the stress in the wall of the tube appears to differ appreciably from the calculated distribution. The experimental technique involved in this type of measurement is however extraordinarily difficult, and this result is not regarded as conclusive. Moreover, the effect if it is real is almost certainly associated with the effect of local yielding of the material under the high stresses set up at the inner corners of the tube.

In addition to the experimental work already noticed, mention must be made of two other experimental methods of analysis. Chitty and Pippard[43] have described a method by which a thin rubber sheet is floated on mercury so that all parts of its surface lie in a true plane, whilst slow motion in the plane is unrestricted. Forces are then applied to the sheet in the horizontal plane, and the resulting displacements are measured microscopically. By this means practically any problem of plane stress or plane strain can be reproduced, and the experimental results for simple cases agree well with the theoretical predictions. Cicala[44] has discussed a method of determining the overall stiffness of structures by observing the interchange of energy between two equal pendulums both suspended from the structure. If the pendulums swing in antiphase, there is no force on the support and the periodic time is proportional to $\sqrt{l}$, where l is the effective length of either pendulum. If they swing in phase, together they exert a force on the structure and the resulting motion of the support increases the period in the ratio $(1+mg/Kl)$, where m is the mass of the pendulum bob and K is the effective stiffness of the structure in the horizontal line in the plane of motion of the pendulums. The beating between these two frequencies causes the amplitude of motion of the two pendulums to wax and wane in antiphase with a period that is directly proportional to the stiffness of the structure. The application of this well-known phenomenon to the actual measurement of stiffness necessitates careful consideration of the effects of the several experimental departures from the simple conditions assumed in the basic

analysis. Thus the behaviour is complicated by the intrinsic damping of the pendulums, by slight differences in their periods and by the dynamic characteristics of the structure. It appears however that with suitable precautions the method can yield reasonably accurate results.

REFERENCES

(1) Black and Southwell. *Proc. Roy. Soc.* A, **164**, 447–67 (1938).
(2) Bradfield and Southwell. *Proc. Roy. Soc.* A, **161**, 155–81 (1937).
(3) Jagn. *Phil. Mag.* **22**, 113–32 (1936).
(4) Broglio. *Ric. di Ing.* **5**, 58–74 (1937).
(5) Shepherd. *Proc. Roy. Soc.* A, **154**, 500–9 (1936).
(6) Seth. *Phil. Mag.* **22**, 582–98 (1936); **23**, 745–57 (1937).
(7) Bradfield, Hooker and Southwell. *Proc. Roy. Soc.* A, **159**, 315–19 (1937).
(8) Filon. *Proc. Roy. Soc.* A, **160**, 137–54 (1937).
(9) Sen. *Phil. Mag.* **24**, 263–72 (1937); *Bull. Calcutta Math. Soc.* **29**, 99–108 (1937).
(10) Fish. *J. Math. Phys.* **14**, 262–73 (1935).
(11) Chitty and Pippard. *Phil. Mag.* **21**, 106–27 (1936).
(12) Tolotti. *R.C. Accad. Lincei*, **25**, 226–30 (1936).
(13) Southwell. *Phil. Mag.* **21**, 201–14 (1936).
(14) Soboleff. *C.R. Acad. Sci., U.S.S.R.*, **4**, 6, 243–6 (1936).
(15) Smirnov. *C.R. Acad. Sci., U.S.S.R.*, **14**, 1, 13–16 and **14**, 2, 69–72 (1937).
(16) Beschkine. *C.R. Acad. Sci., Paris*, **204**, 1918–20 (1937).
(17) Zanaboni. *R.C. Accad. Lincei*, **25**, 117–21 (1937); **25**, 595–601 (1937).
(18) Goodier. *Phil. Mag.* **23**, 186–91 (1937); **23**, 607–9 (1937); **24**, 325–6 (1937).
(19) Southwell. *Proc. Roy. Soc.* A, **154**, 4–21 (1936).
(20) Filon. *Proc. Lond. Math. Soc.* **43**, 152 (1937).
(21) Seth. *Proc. Indian Acad. Sci., Bangalore*, A, **5**, 518–21 (1937).
(22) Miles. *J. Appl. Mech.* A, **3**, 47–52 (1936).
(23) Costello. *Phil. Mag.* **21**, 1005–7 (1936).
(24) Duncan. *Rep. Memor. Aero. Res. Comm., Lond.*, no. 1798 (1937); *Phil. Mag.* **25**, 634–49 (1938).
(25) Frazer, Jones and Skan. *Rep. Memor. Aero. Res. Comm., Lond.*, no. 1799 (1937).
(26) Klotter. *Z. tech. Phys.* **16**, 12, 561–5 (1935).
(27) Tagliacozzo. *Ric. di. Ing.* **4**, 95–8 (1936).
(28) Barta. *C.R. Acad. Sci., Paris*, **204**, 472–3 (1937).
(29) Balinkin. *Phil. Mag.* **21**, 283–90 (1936).
(30) Davies. *Phil. Mag.* **22**, 892–8 (1936); **23**, 464–75 (1937); **23**, 1129–45 (1937).
(31) Pavlik. *Ann. Phys., Lpz.*, **27**, 6, 532–42 (1936).
(32) Procter. *Engineering*, **144**, 62–4, 116–18, 184–6 (1937).
(33) Voinovsky-Krieger. *J. Appl. Mech.* A, **4**, 177–8 (1937).
(34) Leggett. *Proc. Camb. Phil. Soc.* **33**, 325–39 (1937).
(35) Green. *Proc. Roy. Soc.* A, **154**, 430–55 (1936); A, **161**, 197–220 (1937).
(36) Sezawa and Watanabe. *Rep. Aero. Res. Inst. Tokyo*, no. 143, 409–18 (1936).
(37) Titchmarsh. *Sci. Progr.* **91**, 377–86, Jan. 1929.
(38) Avakian. *J. Math. Phys.* **14**, 350–78 (1935).
(39) Pugno. *Atti. Accad. Torino*, **72**, Disp. 1A, 87–95 and 120–7 (1936).
(40) Fairthorne. *Phil. Mag.* **23**, 580–90 (1937).
(41) Southwell. *Proc. Roy. Soc.* A, **163**, 337–55 (1937).
(42) Hovgaard. *Proc. Nat. Acad. Sci., Wash.*, **22**, 362–7 (1936).
(43) Chitty and Pippard. *Proc. Roy. Soc.* A, **156**, 518–37 (1936).
(44) Cicala. *Ric. di Ing.* **4**, 1–8 (1936).
(45) Williams. *Rep. Memor. Aero. Res. Comm., Lond.*, no. 1670 (1935).
(46) Southwell. *Proc. Roy. Soc.* A, **151**, 56 (1935); **153**, 41 (1935).
(47) Bateman. *Phil. Mag.* **19**, 1011 (1935).

IV. THE VISCOSITY OF NON-NEWTONIAN FLUIDS

By C. F. GOODEVE

§ 1. INTRODUCTION

In a previous *Report* Barr[1], in discussing the current theories of viscosity, stated "it may well be that no startling advances will be made until conceptions arising from the study of non-Newtonian liquids have been allowed to influence the general theory of viscosity." The assignment of a fundamental character to the Stokes-Navier relationships between rate of shear and shearing stress (which were based on Newton's hypothesis of motion in liquids) has tended to hide the fact that an enormous number of fluids in common use behave in a manner quite inconsistent with these laws. In the present *Report* the more quantitative or physical aspects of the subject of non-Newtonian fluids are reviewed (other aspects belong more properly to the science of Colloid Chemistry), and the implications of Barr's statement are considered.

The pioneer work on non-Newtonian liquids was carried out by Hatschek[2] and his co-workers, but studies have since been made by a large number of experimenters. The subject has recently been discussed at some length in the First[3] and Second[4] Reports of the Amsterdam Committee on Viscosity, in books by Freundlich[5], Houwink[6], Scott Blair[7] and by the Royal Society[8].

With a Newtonian fluid the applied shearing stress is directly proportional to the rate of flow or of shear, the constant of proportionality being the viscosity. For non-Newtonian fluids the apparent viscosity, i.e. the stress per unit area divided by the rate of shear (velocity gradient), is found to be not only a function of the rate of shear but of the previous history of the sample. A summary of the available knowledge on the subject is only possible, however, if we confine our attention almost exclusively to those experiments in which the effect of the previous history is largely eliminated, i.e. in which a steady state is set up, and to those in which the rate of shear and the shearing stress are known throughout. Only in these cases can the problem be treated with mathematical certainty and only in these cases have the results as yet any physical meaning. This means that we cannot consider any experiments carried out with capillary viscometers, with oscillating cylinders[9], etc. Unfortunately, even in current literature, more than 90 per cent of published work describes such results and unsuccessful attempts at their interpretation. There have, however, been a large number of experiments carried out with viscometers of the coaxial cylinder[10] or cone[11] types in which a fixed amount of fluid is subjected continuously to a nearly uniform shearing motion. It is not suggested that such experiments are of more importance than those which show the effect of previous history, but rather that drastic restriction is necessary if we are to bring order into the very large amount of data which is available.

Such a restriction in scope permits at once a simplified classification and nomenclature, in which the phenomena are divided into three classes:

(1) An isothermal reversible decrease of viscosity with increase in rate of shear. This has been called *thixotropy* (θίξις, the touch; τρέπω, to change) by Freundlich[12, 13], but other terms have been used such as "structural viscosity"[14] or "false body"[9]. In extreme cases substances showing elasticity are liquefied by stirring and gel again when allowed to rest.

(2) An isothermal reversible increase in viscosity with increasing rate of shear. This is *inverse thixotropy*, but in special cases has been called *dilatancy*.

(3) Isothermal irreversible changes.

Many anomalous effects have been attributed to orientation of adsorbed films on the walls[15]. There does not, however, appear to be a clear-cut way by means of which a distinction can be drawn between a wall effect and an anomalous behaviour of class (1) in the fluid itself.

§ 2. THIXOTROPY

The widespread occurrence of thixotropy has only recently been realized. It is most commonly found with the class of fluids known as colloidal solutions. Among the many examples known, most of which are referred to in Freundlich's and Scott Blair's books, the following might be mentioned: paints and inks of almost every type; lacquers and varnishes; greases and graphite lubricants; bitumens; glues and gums; gelatine, casein and starch solutions; liquid cements and concrete; solutions or suspensions of rubber, soaps or dyes; clays of all types; furniture creams and polishes; foodstuffs such as milk, cream, chocolate, etc.; and many medicines and cosmetics. Blood, plant juices and many other biological fluids are thixotropic. The phenomenon is also found with some non-colloidal substances such as lubricating and vegetable oils, especially at high rates of shear and low temperatures.

The flow properties or consistency of the above substances are often of fundamental practical as well as theoretical importance. As examples: (1) A high thixotropy leads to a rapid gelling of a paint immediately after it is brushed and the brush marks are retained. If this is not desired, it is necessary to add substances to lower the thixotropy so that surface tension may have time to level the surface. If the thixotropy is completely destroyed, the paint starts to drain or sag before oxidation hardening takes place. (2) An ideal lubricant from some points of view would be one which has a high viscosity to enable it to be drawn into the bearing, but a much reduced viscosity under high rates of shear so that it offers less resistance and produces a low power loss. Castor and some other oils have this property[16]. (3) Liquid cement sets so rapidly that it generally encloses air bubbles, resulting in a weak product. These may be removed by stirring caused by vibration which lowers the viscosity sufficiently for the bubbles to escape[17].

The most accurate work has been carried out on colloidal solutions synthesized in the laboratory. Almost every insoluble metallic oxide or hydroxide and many insoluble salts form thixotropic colloidal solutions under controlled conditions. Small effects have been found with spherical glass particles[18]. Emulsions of oil in soapy water and foams have been studied and show marked thixotropy

An examination of the colloidal systems exhibiting thixotropy leads to the conclusion that the phenomenon occurs at lower concentrations when the particles are plate- or rod-like. Solid spherical particles only show thixotropy when they aggregate in the form of chains[18, 19] or in the presence of a third species of unsymmetrical shape[20].

The phenomenon is also found to be most marked over a small range of concentrations of certain ions or other precipitating agents. In other words, the thixotropic state is intermediate between the highly dispersed sol-state and that of coagulation. Measurements of the thixotropy yield, therefore, information about the dispersion.

Hydrodynamical considerations. The viscosity of most colloidal solutions containing spherical solid particles conforms with the well-known equation deduced by Einstein[21],

$$\eta = \eta_m (1 + 2{\cdot}5\phi), \qquad \ldots\ldots(1)$$

η being the viscosity of a suspension containing ϕ fraction by volume of spherical particles in a medium of viscosity η_m. It is to be noted that the particles increase the viscous force by acting *through* the dispersion medium. The corresponding equation for elongated ellipsoid particles was deduced from hydrodynamical considerations by Jeffery[22], extended by Eisenschitz, Taylor, Guth and others[23], and confirmed experimentally by Eirich and co-workers[24]. It is now established that if the particles are ellipsoidal or rod-like the factor 2·5 should be very much increased, but, except for the range of rate of shear where there is a transition from slight to almost complete orientation, the viscosity remains Newtonian. J. R. Robinson[25] has recently studied some rather complex systems in this range by simultaneous observations on viscosity and optical rotation. He finds high viscosity to occur with low orientation of the colloidal particles and vice versa, and concludes that the whole of the anomalous viscosity can be explained on hydrodynamic and kinetic grounds assuming free particles. Measurements of temperature coefficients support his conclusions, but the theory is still incomplete.

The most striking thing about the examples quoted in § 2 is that in many cases the increase in viscosity due to the presence of a low concentration of colloidal particles is a hundred or more times greater than that given by the hydrodynamical theory even when account is taken (approximately) of the interference of the particles. Furthermore, the high viscosity of these substances is mainly thixotropic, that is it disappears at high rates of shear.

Quantitative treatment. The resistance to flow shown by a thixotropic fluid in bulk has been treated from two main viewpoints. According to the first the flow rate is plotted against the stress, the latter being considered as the independent variable. Bingham[26] found that the relation between these two quantities was linear over a wide range of stresses but that the line when extrapolated cut the stress-axis at a point some distance from the origin. This point he called the "yield value" and expressed the relation (in other symbols)

$$\sigma = \frac{F - \theta}{\eta_0}, \qquad \ldots\ldots(2)$$

where σ is the rate of shear or velocity gradient (sec^{-1}), F the force per unit area, θ and η_0 constants. In this treatment emphasis is laid upon low rates of shear where thixotropic substances behave as plastic solids.

The other viewpoint is expressed[13, 27] by the equation

$$\eta = \eta_0 + \theta/\sigma \qquad \ldots\ldots(3a)$$

or

$$F = \eta_0\sigma + \theta, \qquad \ldots\ldots(3b)$$

which it is seen is mathematically identical to equation (2), F/σ being the apparent viscosity η. Experimental observations are represented graphically by plotting viscosity against reciprocal shear rate. The rate of shear is now the independent variable but the more important difference between these two viewpoints lies in the fact that in the second the emphasis is laid upon the range of shear rate where the two terms in the equation are comparable, i.e. where the substance behaves as a liquid.

Equations (2) and (3) are applicable to a wide range of substances[13, 26, 27, 28] and can fairly be considered as of general application to thixotropic substances. The first term in equation (3*a*) represents the Newtonian part of the viscosity and usually has a value of the order of magnitude to be expected from hydrodynamical theory. It also is strongly dependent on temperature, a characteristic of Newtonian fluids.

The second term expresses the thixotropic part of the viscosity and θ has been called[13] the "coefficient of thixotropy". In many cases this coefficient is found to be almost independent of temperature, indicating that thixotropy is to be explained by a different mechanism from that of Newtonian viscosity. Not realizing this, many experimenters have expressed their results in units of viscosity relative to that of the solvent. The coefficient of thixotropy is in general the *slope* of the $\eta - 1/\sigma$ graph. It can be measured in dynes/cm^2 and has a well-defined physical significance.

It may be questioned how far measurements made on a large quantity of thixotropic fluid in a coaxial cylinder viscometer describe the behaviour of a typical element of volume of the fluid. Reiner and Riwlin[29], assuming the relation given in equation (3) for an element of volume, have arrived at an integrated equation for the Couette apparatus. Goodeve and Stevenson[30] have further shown that, for layers of liquid thin compared with the radius of the cylinders, the observed apparent viscosity is identical with the true elementary viscosity. Richardson and Tyler [31] have measured directly the flow from point to point in a coaxial cylinder viscometer, and obtained results in conformity with equation (3). Therefore for substances whose viscosity follows equation (3), the quantitative treatment is well founded.

Theories of thixotropy. Extensive experiments carried out by Ostwald and coworkers[14] led them to conclude that there are structures present in non-Newtonian fluids. Bingham[26] considers that in the case of concentrated suspensions, such as paints and pastes, the particles are in actual contact with one another, forming a bridge-like structure throughout the bulk phase. He considers that these particles

slide over one another and that in this motion they obey the ordinary laws of friction, that is the force is independent of the velocity, as in the second term of equation ($3b$). He considers that this force remains constant down to zero rates of shear, where it becomes the yield value.

Bingham's viewpoint has been criticized by Hatschek[32] and others. Goodeve and Whitfield[13] pointed out that, although the formation of gels and the anomalies in viscosity arose from the same cause, there could in most cases be no quantitative agreement between the observed yield value and the value of θ, as the structure responsible for the elasticity of the gel is to some extent broken down when the system is moving as a fluid under a uniform rate of shear. They put forward an alternative theory in which account was taken of the rate of building up of the structure due to Brownian motion.

In addition to the hydrodynamical theory referred to above, other theories have been put forward, some introducing the elasticity of the particles[33], others[34] being developments of the original relaxation theory of Maxwell for the flow of metals under stress. Eyring[35], in the course of an interesting treatment of Newtonian viscosity according to the "transition state method" of reaction kinetics, pointed out that colloidal particles, by virtue of their size, are able to concentrate at a point large amounts of energy derived from the shear. The "impulse theory" recently brought forward appears to have fairly wide application. Before considering this, however, it would be well to refer to the various types of structure which have been postulated.

Types of structure. From the fact that most colloidal suspensions when allowed to sediment occupy a volume very much in excess of that to be expected if there were close packing, many experimenters have drawn the conclusion that colloidal particles are surrounded by an immovable layer of solvent, often as much as 100 molecules thick[36, 37]. Despite the fact that the existence of these "lyospheres" has never been shown experimentally and that they are not to be expected in view of our present knowledge of the structure of liquids[38], this idea has persisted in many quarters[39]. It would appear that such solvent as may be bound[40] plays an insignificant part in the properties of the solution.

In some quarters[5, 41] this theory has largely been replaced by another in which long range attractive and repulsive forces are assumed to exist, of a magnitude such that they can give rise to a potential energy minimum lying at some distance from the surface of each particle. Colloid particles would tend to lie in the position of these minima and to form a system of close packing which resists distortion. While there are theoretical[42] arguments in favour of the existence of these forces, there is no experimental evidence[43]. It has been shown[27] that this theory cannot account for the elasticity of gels made up of rod-like particles—one of the most important classes of gels with which we have to deal.

On the other hand a structure in which the particles are actually in contact with one another, forming a scaffolding from wall to wall, has been shown by Usher and co-workers[19, 44] and others[27] to account very satisfactorily for all of the observations including the high sedimentation volume. In particular it might be pointed

out that McDowell and Usher[44] showed that gels of carbon black in non-conducting mixtures of organic solvents conducted the electric current when allowed to stand but ceased to do so when the liquid was stirred. Hauser and Reed[45] have observed under the ultramicroscope, the formation of a scaffolding structure in bentonite suspensions.

The impulse theory of thixotropy[27]. As described above, the thixotropic part of the viscosity appears to act by a mechanism independent of that of the Newtonian part. There appears to be no doubt about the fact that when a liquid is sheared the structure is broken down. At the same time, however, this structure is being rebuilt by the particles being brought into contact through their Brownian motion *and* by the action of the shearing motion itself. The links between the particles are continually being stretched until they are broken, reforming and being stretched again. During this stretching process an impulse is passed from a fast moving layer to a slower moving layer. If we sum up the total number of impulses per second between any two layers we arrive at the total force acting between them. It has been shown that if we assume Hook's law to apply to the stretching process, the value of each impulse is *inversely* proportional to the rate of shear. This arises from the fact that if the shear rate is increased the time of stretching is decreased, the force rising to the same critical value where the link breaks. As the number of impulses is proportional to the rate of shear, a final equation is obtained

$$F = \tfrac{1}{2} n z f_c. \qquad \text{.....(4)}$$

In this equation f_c is the critical force which the link will withstand, z the distance between the layers (related to the particle size), and n the number of links per c.c. The force is therefore equal to the number of links per cm^2 of surface between the layers, multiplied by the average force $\frac{1}{2} f_c$, as is to be expected. z and f_c are effectively constant and in the range of shear rate where the rate of link formation due to the shear is greater than that due to the Brownian motion, the value of n is also constant. We should expect therefore that the increase in viscosity, due to the presence of colloidal particles, the links between which are stretched and broken by the shear, should be a constant independent of the rate of shear and of the temperature. These are the characteristics found for thixotropic systems.

The impulse theory gives an explanation of the difference between thixotropic and Newtonian behaviour. If the energy required to break a link is not too great compared with kT, the life of the link will be controlled entirely by the thermal energy rather than by the rate of shear and will be a constant at constant temperature. The stretching process now lasts for only a fixed time and in this time can develop a force and give an impulse *proportional* to the rate of shear. The total force also becomes proportional to the rate of shear as shown by the following equations,

$$F = I \frac{nz}{\tau} = \tfrac{1}{2} H z \sigma \tau^2 \frac{nz}{\tau} = \tfrac{1}{2} n H z^2 \sigma \frac{1}{\nu} e^{E/kT}. \qquad \text{.....(5)}$$

I is the impulse, τ the mean life, equal to $\frac{1}{\nu} e^{E/kT}$, H the Hook's modulus, ν the

frequency of interchange of energy and E the energy necessary to break the link. In addition to describing the behaviour associated with Newtonian fluids this equation gives the well-known exponential relation between the viscosity coefficient and the temperature.

Summarizing, we see that the chief distinction between thixotropic and Newtonian behaviour can be attributed to the possibility of the shear influencing the average life of a link. In colloidal solutions we have links of high energy compared with kT, but there are relatively few of these links per c.c. and, by virtue of the magnitude of the colloidal particles, energy derived from the shear may be concentrated in such a way that it breaks the links. In a glass (which is homogeneous), on the other hand, there are a large number of links of great strength and at the same time there is insufficient structure by means of which energy can be concentrated. The high Newtonian viscosity of glass follows therefore from equation (5). Ordinary liquids have a large number of links of low energy and thus exhibit low Newtonian viscosity.

Lubricating and vegetable oils form an intermediate class. Their long irregular molecules form "links" merely by entanglement; these when sheared will be stretched and broken as described. At low rates of shear, disentanglement takes place due to thermal energy and the liquids are Newtonian. At high rates of shear, the disentanglement is assisted by the shearing motion, and we get the conditions leading to thixotropic behaviour.

Foams and emulsions do not have links between the bubbles or droplets but resist flow at high concentrations by virtue of the fact that they become distorted—their surfaces increase against the force of surface tension. A treatment of this type of system[27] leads to the conclusion that the viscosity should be partly Newtonian and partly thixotropic, as is found in practice.

Conclusions. The impulse theory leads to a qualitative description of the part played by the previous history in the anomalous behaviour of the substance. For example, in the absence of shearing motion the value of n (equation (4)) should increase at a rate which increases with temperature and with decrease in particle size. This is observed experimentally (ref.[46] and[43] resp.). Now that it is possible to describe accurately the steady state of a thixotropic fluid undergoing shear, it may soon become possible to describe the effect of the previous history equally accurately.

There are however a number of cases in which the value of n is in no way a constant but decreases markedly with increase of shear rate. Modifications in the theory will have to be made to take account of these cases.

§ 3. INVERSE THIXOTROPY

Cases in which the viscosity *increases* reversibly with increasing rate of shear are not so common. There is of course the classical example of wet sand, which when walked upon appears to become dry. This drying is due to the fact that the sand granules are normally in a state of close packing and if sheared must separate

before any one particle can slide over another. This has been called dilatancy by Osborne Reynolds[47]. The important fact that quicksand may be distinguished from ordinary sand, in that the former is thixotropic and not dilatant, has been pointed out by Freundlich and Juliusburger[48].

Other examples have been investigated quantitatively by Freundlich and Röder[49], who found that dilatancy was observed only over a narrow range of concentration. The phenomenon is very marked with a suspension of 43 per cent rice starch in water; the rate with which a ball can be pulled out of a suspension (by means of a wire) is independent of the applied force over a wide range. While dilatancy is generally associated with close packing, these authors point out that 43 per cent corresponds to a fairly loose packing and suggest a piling up effect.

Without going into details, it is easy to see why concentrated suspensions of rigid particles should jam when an attempt is made to shear them quickly. Movement, however, is not completely stopped, presumably because the particles will have thin "lubricating" layers of liquid between them. This layer will decrease in thickness as the shearing stress is increased and therefore there will be an increase in the resistance to motion of the whole system. Such layers must exist, otherwise the particles would come into contact and, if incompressible, as for example in quartz, fluid motion would be impossible.

It is of interest to note that Freundlich and co-workers found that quartz and starch particles show dilatancy in polar liquids but thixotropy in non-polar liquids. Presumably in the latter case the liquid does not prevent the formation of links between the particles. A very full discussion of the whole question has been made in a paper by Freundlich and Jones[50].

Hatschek and Jane[51] observed that dilute ammonium oleate solutions apparently increased their viscosity with increased rate of shear. This phenomenon was subsequently[52] attributed to the presence of turbulent flow. Lawrence[53] developed a beautiful method to follow the flow of fluids in a tube and found that the viscosity of ammonium oleate solutions increases with shear at low rates of shear until it reaches a maximum beyond which it decreases, i.e. becomes thixotropic. More recently Hartley[54] has shown that a true increase of viscosity with increasing rate of shear most certainly takes place with a number of soap-like compounds. The phenomenon has been observed down to concentrations of 0·02 per cent, and viscosity increases of tenfold to one hundredfold have been observed. While this phenomenon has only been studied in a qualitative manner and much more evidence will be necessary before it can be explained, it appears likely that the molecules or particles are coiled up or are in the well-known micelle form at low rates of shear, but are straightened or stretched out to great lengths at the higher rates. According to the hydrodynamical theory[23] a particle contributes to the viscosity according to its length raised to a power between one and two.

Liquid helium II is apparently anomalous; its rate of flow has been found[55] to be independent of the pressure. There is no obvious relation between this non-Newtonian behaviour and the others considered above, but an explanation will most likely be of value to the general theory of viscosity.

§ 4. IRREVERSIBLE EFFECTS

The question of reversibility involves the time factor, and many of the examples discussed in the preceding two sections show marked hysteresis, particularly at low rates of shear. This is in general due to the fact that the Brownian motion responsible for building up structures in this range is small. For example, a paint whose viscosity has been measured at a low shear rate may require several days to regain the same viscosity after it has been subjected to a very high rate of shear for a short period. With many other examples, however, hysteresis is undetectable with the instruments available. Classifications, based on the amount of the hysteresis, have been put forward[7,9], but as they depend on the instrument and on the observer they are unlikely to receive general acceptance.

On the other hand, most protein solutions when subjected to a very high rate of shear undergo a permanent decrease in viscosity. This is presumably due to an irreversible chemical change arising from the breaking of the chain-like molecules. This phenomenon has also been found with many synthetic highly polymerized substances. Milk[56], on the other hand, increases in viscosity after subjection to high shear rates.

§ 5. NEWTONIAN VISCOSITY

Although most pure liquids are rigidly Newtonian, no completely satisfactory treatment of this behaviour has been put forward. Many theories[3,4] have developed out of the original ideas of Maxwell and of van der Waals, Jnr., and each has its particular range of application. For simple Newtonian liquids, the impulse theory is probably inferior to many of the others but has some advantages for associated liquids such as glycerine, treacle, etc. It may well be that these liquids would show thixotropic behaviour if subjected to high enough rates of shear.

REFERENCES

(1) Barr, G. *Report on Progress of Physics*, **3**, 23 (1936).
(2) Hatschek, E. *The Viscosity of Liquids* (Bell and Sons, 1928).
(3) *First Report on Viscosity and Plasticity*, by a Committee of the Academy of Sciences at Amsterdam (1935).
(4) *Second Report on Viscosity and Plasticity*, by a Committee of the Academy of Sciences at Amsterdam (1938).
(5) Freundlich, H. *Thixotropy* (Hermann, Paris, 1935).
(6) Houwink, R. *Elasticity, Plasticity and Structure of Matter* (Cambridge, 1937).
(7) Scott Blair, G. W. *An Introduction to Industrial Rheology* (Churchill, London, 1938).
(8) Lawrence, A. S. C., Hatschek, E., Schofield, R. K. and others. *Proc. Roy. Soc.* A, **163**, 319 (1937).
(9) Pryce-Jones, J. *J. Oil Col. Chem. Ass.* **17**, 305 (1934); **19**, 295 (1936).
(10) See Barr, G. *Monograph on Viscometry* (1931), or Hatschek, ref. (2).
(11) Goodeve, C. F. *J. Sci. Instrum.* **16**, (1939) in proof, and ref. (13).
(12) Freundlich, H. e.g. *J. Phys. Chem.* **41**, 901 (1937).
(13) Goodeve, C. F. and Whitfield, G. W. *Trans. Faraday Soc.* **34**, 511 (1938).
(14) Ostwald and others. Many papers in *Kolloidzschr.* since 1925.

(15) These anomalies have been dealt with fully by JAEGER, F. M. (see ref. (4), p. 100).
(16) NEALE, S. M. *Chem. and Ind.* **56**, 140 (1937).
(17) CASSIE, W. F. *Struct. Engr*, **13**, 326 (1935).
(18) BROUGHTON, G. and WINDEBANK, C. S. *Industr. Engng Chem.* **30**, 407 (1938).
(19) USHER, F. L. *Proc. Roy. Soc.* A, **125**, 143 (1929).
(20) RUSSELL, J. L. and RIDEAL, E. K. *Proc. Roy. Soc.* A, **154**, 540 (1936).
(21) EINSTEIN, A. *Ann. Phys., Lpz.*, **19**, 289 (1906); **34**, 591 (1911); see also ref. (4), p. 128.
(22) JEFFERY, G. B. *Proc. Roy. Soc.* A, **102**, 161 (1922).
(23) A full treatment is given by BURGERS, J. M. ref. (4), p. 134.
(24) EIRICH, F., MARGARETHA, H. and BUNZL, M. *Kolloidzschr.* **75**, 20 (1936).
(25) ROBINSON, J. R. Private communication, a thesis submitted to the University of Cambridge, 1938.
(26) BINGHAM, E. C. *Fluidity and Plasticity* (McGraw Hill, 1922).
(27) GOODEVE, C. F. *Trans. Faraday Soc.* in the press.
(28) ROSSI, C. *Gazz. chim. ital.* **67**, 691 and 751 (1937); **68**, 3 (1938).
(29) REINER, M. and RIWLIN, R. *Kolloidzschr.* **43**, 1 (1927).
(30) GOODEVE, C. F. and STEVENSON, A. C. Unpublished calculations (see ref. (11)).
(31) RICHARDSON, E. G. and TYLER, E. *Proc. Phys. Soc.* **45**, 142 (1933).
(32) Ref. (2), p. 209.
(33) SZEGVARI, A. *Z. phys. Chem.* **108**, 175 (1924).
REINER, M. *Kolloidzschr.* **39**, 80 (1926).
(34) See ref. (7), Chapter v.
(35) EYRING, H. *J. Chem. Phys.* **4**, 283 (1936).
(36) HAUSER, E. A. *J. Rheology*, **2**, 5 (1931).
(37) GESSNER, H. *Kolloidchem. Beih.* **19**, 213 (1924).
(38) See Faraday Society General Discussion. *Trans. Faraday Soc.* **33**, 1 et seq. (1937), in particular a paper by BERNAL, J. D.
(39) E.G. VON BUZÁGH, A. *Colloid Systems* (Technical Press, London, 1937).
(40) HATSCHEK, E. *Trans. Faraday Soc.* **32**, 787 (1936).
(41) HAMAKER, H. C. *Rec. Trav. chim. Pays-Bas*, **56**, 727 (1937), etc.
(42) KALLMANN, H. and WILLSTAETTER, M. *Naturwissenschaften*, **20**, 952 (1932).
(43) BASTOW, S. H. and BOWDEN, F. P. *Proc. Roy. Soc.* A, **151**, 220 (1935); *Nature, Lond.*, **135**, 828 (1935).
(44) MCDOWELL, C. M. and USHER, F. L. *Proc. Roy. Soc.* A, **131**, 409 and 564 (1931).
(45) HAUSER, E. A. and REED, C. E. *J. Phys. Chem.* **41**, 911 (1937).
(46) BROUGHTON, G. and SQUIRES, L. *J. Phys. Chem.* **40**, 1041 (1936).
(47) REYNOLDS, OSBORNE. *Phil. Mag.* (5), **20**, 469 (1885).
(48) FREUNDLICH, H. and JULIUSBURGER, J. *Trans. Faraday Soc.* **31**, 769 (1935).
(49) FREUNDLICH, H. and RÖDER, H. L. *Trans. Faraday Soc.* **34**, 308 (1938).
(50) FREUNDLICH, H. and JONES, A. D. *J. Phys. Chem.* **40**, 1217 (1936).
(51) HATSCHEK, E. and JANE, R. S. *Kolloidzschr.* **38**, 33 (1926).
(52) ANDRADE, E. N. DA C. and LEWIS, J. W. *Kolloidzschr.* **38**, 260 (1926).
(53) LAWRENCE, A. S. C. *Proc. Roy. Soc.* A, **148**, 59 (1935).
(54) HARTLEY, G. S. *Nature, Lond.*, **142**, 161 (1938).
(55) ALLEN, J. F. and MISENER, A. D. *Nature, Lond.*, **142**, 643 (1938).
(56) BUGLIA, C. *Kolloidzschr.* **2**, 353 (1908).

THE ADSORPTION THEOREM OF J. WILLARD GIBBS. ANOMALIES IN APPLYING IT TO SURFACE TENSION CURVES THAT EXHIBIT MINIMA IN DILUTE SOLUTION

By J. W. McBAIN, F.R.S. and G. F. MILLS
Stanford University, California

§ 1. INTRODUCTION

The thermodynamic relation of J. Willard Gibbs between surface tension and adsorption is universally taught as the very foundation of surface phenomena. Nevertheless, this theorem occupies the unique position in science of a thermodynamic relation which commands belief but yet is commonly held to be inapplicable to a large group of instances because its consequences would be inacceptable. The reasons given by numerous authors for not applying it while paying it lip service really amount to a denial of the validity of thermodynamics.

First, we may recall the nature of the Gibbs relation. Stated generally, if the surface tension of any binary solution is plotted against the logarithm of the activity of the solute, the theorem demands that adsorption must depend solely upon the slope at any given point, regardless of the nature of the solute, and also, surprisingly, regardless of the magnitude of the surface tension. For example, it is a rigorous requirement that if, at a given point or concentration, the surface tension curve happens to be horizontal, there can be no adsorption; that is, no difference whatsoever in composition of the surface layers as compared with the bulk of the solution. This is required no matter whether the surface tension at this point is equal to that of the solvent, is greater, or lies far below that of the solvent.

Now, as will be shown, there are many cases, especially with modern detergents, where such horizontal portions occur in extreme dilution. Here the surface tension is only a fraction of that of pure water and yet the Gibbs equation states that there is no adsorption to account for the enormous lowering of surface tension!

Elsewhere we have classified as type III, surface tension curves exhibiting great lowering in extreme dilution, the surface tension thereafter remaining at the constant low value with increasing concentration or after passing through a minimum rising to a low maximum while still in dilute solution and then diminishing again. There are about one hundred known examples which have been obtained by practically all the known methods of measuring surface tension. The references to these are collected in the appendix at the end of this paper. Type III curves are thus very different from type I curves where there is a progressive lowering of surface tension with concentration, and from type II curves where it is progressively though slightly raised as is the case for solutions of many electrolytes and sugar.

The viewpoint of the following discussion is to insist that thermodynamics must be applicable in all cases whilst pointing out the impossible conclusions arising from the direct application of the Gibbs theorem. The conclusion is therefore that Gibbs's treatment is incomplete, having left out important factors.

Excellent, critical, theoretical modern discussions of the derivation and interpretation of the Gibbs theorem have recently been given by Guggenheim and Adam[1], by Koenig[2], by Defay[3], and by Gatty[4] and his collaborators. These should be compared, but need not here be repeated. We must quote however the two original equations of Gibbs's[5] for a two-component solution:

$$\delta\epsilon^s = t\,d\eta^s + \sigma\delta s + \mu_1\delta m_1^s + \mu_2\delta m_2^s \qquad \text{.....(497)}$$

and

$$d\sigma = -\eta_s\,dt - \Gamma_1 d\mu_1 - \Gamma_2 d\mu_2, \qquad \text{.....(508)}$$

where the superscript or subscript s indicates that the quantities relate to the surface, σ is the surface tension, μ_1 and μ_2 the chemical potentials of the two components ($d\mu = RT \times d \ln a$, where $\ln a$ is the natural logarithm of the activity), ϵ the internal energy, t the temperature, η the entropy, s the area of surface, m_1 and m_2 amounts of constituents, and finally Γ_1 and Γ_2 the adsorptions of solvent and solute. Γ_1 and Γ_2 are measured as differences in amount per square centimetre of surface when an arbitrary but inclusive thickness of the surface layer is compared with the composition of the homogeneous solution.

The Gibbs theorem is usually applied in the following form to a solution at constant temperature:

$$\Gamma_2' = -\frac{d\sigma}{d\mu} = \frac{1}{RT}\frac{d\sigma}{d\log_e a}, \qquad \text{.....(A)}$$

where Γ_2' is the adsorption or excess of the solute in the surface, the boundary being so chosen that the value of Γ_1 for the solvent is made to vanish.

§ 2. THE NUMERICAL USE OF CHEMICAL POTENTIAL OR ACTIVITY IN THE GIBBS EQUATION

The actual use of chemical potentials or activity has been rare because it is too often supposed that it is necessary to determine activity coefficients, requiring a somewhat elaborate study of solutions down to infinite dilution. Hence usually concentration c is substituted for activity a as a convenient approximation in equation (A). We shall now point out that it is really very simple to employ correct thermodynamic values without elaborate experimentation because for this purpose a knowledge of the actual value of the activity coefficients is wholly unnecessary.

However, we may digress for a moment to point out that even when concentration is used a minimum in the surface tension curve would equally exclude adsorption. Thermodynamic theory demands, for any stable solution that does not spontaneously break up into separate parts, that chemical potential and activity must show positive increase with increasing concentrations. Hence nothing is altered in the sign but only in the amount by the predictions of the theorem when concentration is used instead of activity.

It will be noted that in the thermodynamic equation, not the activity but only the change in its logarithm enters into the equation. Hence the result is wholly independent of any units or standard value assigned to the activity in any one solution. The osmotic coefficient g is mathematically identical with Lewis and Randall's $1-j$ and equals van't Hoff's i/ν. The lowering of freezing point with modern commercial equipment requires less than 10 min. for measurement to 0·01° and directly gives $g=\theta/\nu\lambda m$. Only a few measurements in the immediate range of concentration under examination are required for thermodynamic purposes.

It may readily be shown in a manner similar to that used by Bjerrum in deriving the differential relation between osmotic and activity coefficients that

$$\frac{d\log_e a}{dm_2}=\nu\frac{dg}{dm_2}+\nu\frac{g}{m_2}=\frac{di}{dm_2}+\frac{i}{m_2},$$

where ν is unity for a non-electrolyte and is equal to the number of ions at infinite dilution for an electrolyte, and natural logarithms ln to the base e are used. The expression $d\ln a$ is accurately defined as $d\log_e(N\pm f\pm)^\nu$, according to equations (19) and (21) of van Rysselberghe[6].

If the solution is semi-ideal (Guggenheim), this will be true at room temperature as well as at the freezing point. Hence Γ is immediately calculable from g together with surface tension, as will be done in table 1. First, the surface tension is plotted against m_2, the slope giving $d\sigma/dm_2$. Next, g is plotted against m_2, yielding the two terms on the right of our equation and therefore $d\log_e a/dm_2$. The quotient $(d\sigma/dm_2)/(d\log_e a/dm_2)=d\sigma/d\log_e a=-RT\Gamma_2'$. For a different choice of the Gibbs surface in very dilute solutions $\Gamma_2-(m_2^\alpha/m_1^\alpha)\Gamma_1$ is identical with Γ_2', illustrating the fact that the adsorption is a property of the system and is not actually altered by different choices for the position of the hypothetical mathematical surface of Gibbs's. It will be noted that activity, like activity coefficient, is neither evaluated nor required for the purpose of the exact Gibbs theorem $\Gamma_2'=-d\sigma/RTd\log_e a=-d\sigma/d\mu$. All that is necessary is the numerical ratio of the activities of successive concentrations of solution in the immediate range under study. Indeed, if g is changing but little with concentration at the moment, an excellent approximation to the exact value of Gibbs's adsorption Γ_2' is obtainable from a single measurement of freezing point by using $\nu gd\ln c$ or $id\ln c$ instead of $d\ln a$, a simplification first suggested and used by Milner[7] in 1907.

§ 3. THE RESULT OF APPLYING THE CLASSICAL GIBBS THEOREM TO TYPE III CURVES OF SURFACE TENSION

Of the numerous type III curves (appendix) only two examples will be chosen for calculation. The significance of these type III curves arises from the surface tensions being very greatly lowered in extreme dilution, so that where the Gibbs theorem demands that the surface and solution be of identical composition, the surface, like the solution, is nearly pure solvent. As will appear, there is no such interest in the examination of certain concentrated binary mixtures even where

there is sometimes a lower surface tension than that of either component, because this still permits of a monolayer of the two components in comparable proportions, easily accounting for the observed surface tension.

The surface tension curve of lauryl sulphonic acid, $C_{12}H_{25}SO_3H$, in aqueous solution, as obtained in the Stanford laboratory with the du Noüy tensiometer, exhibits a minimum at a concentration of 0·006 mol. (or 99·85 per cent water). This portion of the surface tension curve is given in figure 1. The corrections suggested by Harkins and Jordan were taken into account.

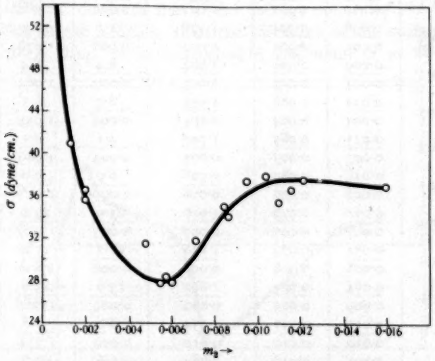


Figure 1. Surface tension of lauryl sulphonic acid at 22° C.

The adsorption of the sulphonic acid was calculated from the slopes and concentrations of figure 1 by the classical Gibbs theorem, using the unpublished freezing point data of Dr K. Andresen to calculate the activities or thermodynamic potentials required. Details of the calculations are recorded in table 1, the procedure and nomenclature having been described in the preceding section.

Guggenheim uses the notation Γ_2' to denote the usual convention that the adsorption is measured solely with reference to amount of water present; that is, that the adsorption of the water taken as is zero. Γ_2' is used in figure 2 to express the amount of adsorption of lauryl sulphonic acid in mol./cm.^{2} so deduced from the slope of surface tension with chemical potential.

The remarkable features arising from the use of Gibbs's theorem in this particular case are that the adsorption increases rapidly from zero at infinite dilution to the value $4{\cdot}0 \times 10^{-10}$ mol./cm.^{2}, for $m_2 = 0{\cdot}0013$, falling as expected to zero at 0·006, proceeding to a minimum negative value of $-5{\cdot}09 \times 10^{-10}$ mol./cm.^{2} at 0·007, and rapidly rising again through the value zero at about 0·012 m_2, still in the very dilute region, and then going on to a second region of positive adsorption.

Table 1. The calculation of the adsorption of lauryl sulphonic acid in the surface of its solutions at 22° c., by the classical adsorption theorem of Gibbs's, in mol./cm.²

$$\Gamma_2' = -\frac{d\sigma}{d\mu} = -\frac{1}{RT}\frac{d\sigma}{d\log_e a} = -d\sigma/(\nu dg + \nu g\, d\log_e m_2) = -d\sigma/di + i\, d\log_e m_2)$$

m_2	σ^*	$\frac{-dg}{dm_2}$	g/m_2	$\nu\left(\frac{g}{m_2}-\frac{dg}{dm_2}\right)$	$\frac{-d\sigma}{dm_2}$	$RT\Gamma_2'$	$\Gamma_2' \times 10^{10}$†
0·001	45·7	0·003 / 0·001	0·997 / 0·001	1·988 / 0·001	16·0 / 0·001	16·0 / 1·988	3·30
0·002	36·0	0·006 / 0·002	0·994 / 0·002	1·976 / 0·002	11·0 / 0·002	11·0 / 1·976	2·28
0·003	32·1	0·009 / 0·003	0·991 / 0·003	1·964 / 0·003	8·4 / 0·003	8·4 / 1·964	1·755
0·004	29·5	0·012 / 0·004	0·988 / 0·004	1·952 / 0·004	6·5 / 0·004	6·5 / 1·952	1·365
0·005	28·2	0·015 / 0·005	0·985 / 0·005	1·940 / 0·005	4·1 / 0·005	4·1 / 1·940	0·85
0·006	28·1	0·018 / 0·006	0·982 / 0·006	1·928 / 0·006	0·0 / 0·006	0·0 / 1·928	0·00
0·007	30·4	0·021 / 0·007	0·979 / 0·007	1·916 / 0·007	−23·9 / 0·007	−23·9 / 1·916	−5·09
0·008	33·2	0·054 / 0·008	0·973 / 0·008	1·838 / 0·008	−17·2 / 0·008	−17·2 / 1·838	−3·84
0·009	35·4	0·064 / 0·009	0·965 / 0·009	1·802 / 0·009	−13·6 / 0·009	−13·6 / 1·802	−3·10
0·010	36·8	0·090 / 0·010	0·957 / 0·010	1·734 / 0·010	−8·8 / 0·010	−8·8 / 1·734	−2·16
0·011	37·6	0·105 / 0·011	0·947 / 0·011	1·684 / 0·011	−5·6 / 0·011	−5·6 / 1·684	−1·365
0·012	37·7	0·120 / 0·012	0·937 / 0·012	1·634 / 0·012	−1·1 / 0·012	−1·1 / 0·012	−0·276
0·013	37·5	0·155 / 0·013	0·926 / 0·013	1·542 / 0·013	2·9 / 0·013	2·9 / 1·542	0·766
0·014	37·3	0·214 / 0·014	0·914 / 0·014	1·400 / 0·014	3·1 / 0·014	3·1 / 1·400	0·90
0·015	37·1	0·240 / 0·015	0·900 / 0·015	1·320 / 0·015	3·3 / 0·015	3·3 / 1·320	1·02

* From smooth curve. † $RT = 2{\cdot}45 \times 10^{10}$.

Two most conspicuous features common to all type III curves possessing a minimum (i.e. $d\sigma/d\mu = 0$) are: First, the classical thermodynamic equation requires that there be no adsorption whatsoever where the surface tension is most greatly lowered, leaving no mechanism for lowering it. In the present case, the surface tension of water is lowered to less than half value, 28·1 dynes, where the theorem denies any effects of adsorption. Secondly, that while the solution is still very dilute and the surface tension only a fraction that of water, a very large negative adsorption of the solute is demanded. The physical significance of the latter must of course be an enormous positive adsorption of the solvent in this region of very

low surface tension, for $m_1 d\mu_1 + m_2 d\mu_2 = 0$ and therefore $d\sigma/d\mu_2 = \Gamma_1 m_2/m_1 - \Gamma_2$. An identical numerical result follows from placing the Gibbs surface so that the adsorption of the solute is equated to zero, giving the positive adsorption of pure solvent $\Gamma_1^{2} = \Gamma_2' m_1/m_2 = 40{,}400 \times 10^{-10}$ mol. of water/cm^2; that is, the depth of the layer of pure water on the surface of the solution comes out from the Gibbs equation as 7860 A. deep, where the surface tension is so greatly lowered! If it be supposed that any solute is anywhere near the surface, this thickness is proportionally increased. For every molecule of solute assumed put into the surface layer 11,500 molecules of water must be added also to satisfy the Gibbs equation. For example, the theorem demands that if the average concentration of the surface layer is half that of the solution, then the thickness of the surface is doubled, becoming 15,700 A. deep. It is inconceivable that water or so dilute a solution

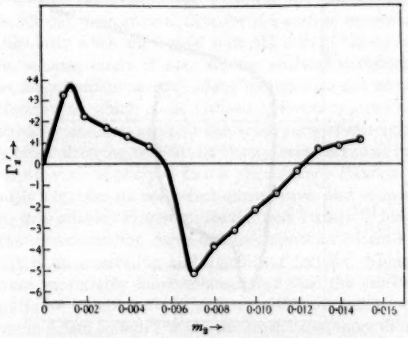


Figure 2. Adsorption of lauryl sulphonic acid at 22° C.

monopolizing the whole depth of the surface layer can so greatly lower the surface tension of water.

The really unaltered surface tension of these very dilute solutions as measured by a sufficiently rapid dynamic method before any diffusion or rearrangement could occur should be indistinguishable from that of water (cf. Bond and Puls[8]). Only a high degree of adsorption of sulphonic acid in all these cases can explain the very great lowering of the surface tension to the static value. This is borne out by the fact that all these solutions froth freely, a proof according to Foulk[9] that the surface is of composition different from that of the bulk of the solution, requiring osmotic work to be performed if the attempt is made to equalize the two.

Again, the theorem is at variance with the absolute magnitude of the surface tension in that, as has just been shown in figures 1 and 2, when the adsorption is again supposedly positive, the surface tension is higher than where no adsorption at all occurs or where, in slightly less dilution, the surface is deeply covered with

water. Indeed, as is shown in figure 3, the surface tensions over a whole range have the same value whether the supposed adsorption is deduced by the Gibbs theorem as positive or negative, solute or solvent. If both adsorptions of solute and solvent are calculated as positive, excess or deficiency of solute in the surface solution leads equally to the same surface tension. The whole situation is impossible if the Gibbs theorem is applied in its accurate classical form.

Direct measurements in the Stanford laboratory with the microtome method[10] have established that over the whole concentration range here discussed positive

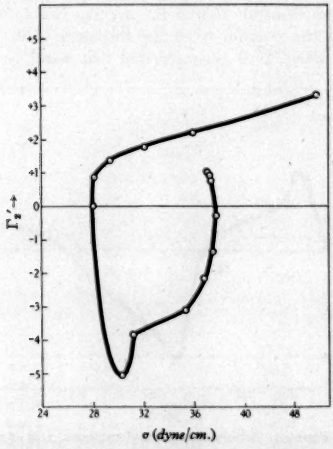


Figure 3. Adsorption as calculated from Gibbs's theorem, showing no relation with absolute lowering of surface tension, excess or deficiency of solute in the surface solution producing the same surface tension, for type III curves exhibiting a minimum.

values for adsorption of lauryl sulphonic acid are actually obtained in striking contrast to the predicted values in table 1 and figures 2 and 3. These will be communicated in full elsewhere.

Quite apart from type III curves, Gibbs's theorem shows for some of the simplest types of surface tension curve that the surface tension is not determined solely by the outermost layer of molecules. In these well-known simple cases (Szyszkowski, Langmuir), such as aqueous solutions of the soluble lower fatty acids, the surface tension decreases linearly with the logarithm of the concentration, and therefore in spite of $\Gamma_2' = d\sigma/d \ln c$ yielding a constant monomolecular layer of adsorbed solute the surface tension is progressively lowered as the underlying submerged substrate increases in concentration.

A second example to which we may refer more briefly is the surface tension of

very dilute solutions of ordinary binary electrolytes, using the precision data of Grinnell Jones and Ray[11] for sodium chloride as a typical illustration. In 0·001 molar solution the surface tension is slightly lowered to a minimum value. The Gibbs equation, on the assumption that this is a two-component system, requires that the adsorption is zero. In slightly less dilute solution, the surface tension is rising and the adsorption is negative corresponding at 0·002 molar to a layer of pure water 8 Å. or several molecules deep; or proportionately thicker if there is any trace of salt so near the surface.

§ 4. REASONS WHICH HAVE BEEN SUGGESTED FOR NOT APPLYING GIBBS'S THEOREM

The most obvious suggestion for avoiding the application of the Gibbs theorem is that all the experimental data for the surface tension of solutions are erroneous, particularly when they yield type III curves. There is some merit in this contention because rarely if ever are the surfaces demonstrably clean and maintained free from contamination. Many methods do not allow sufficient time for final equilibrium, to which alone Gibbs's theorem is strictly applicable, and many are affected by adsorption upon the solid parts of the equipment, contact angle probably often changing without its change being noted or measured. This is probably the explanation of the von Lerch phenomenon (Harkins[12]). Thus drop number is usually regarded as only semi-quantitative, and drop weights are only somewhat more dependable. Edgerton, Hauser and Tucker[13] have shown that the minimum in the surface tension curve becomes apparent within a few hundredths of a second. It is an interesting and significant fact (cf. Milner[7]) that in all solutions that are appreciably more concentrated than the minimum, the surface tension appears to be instantaneously and reversibly arrived at, as studied with a Pockels-Langmuir-Adam trough[13]. On the other hand, more dilute solutions may require several days for the surface tension to be fully lowered. This is particularly striking where the theorem demands that there is no adsorption and no diffusion, or where it is supposed that water is accumulating over the surface.

Despite the fact that every known method of measuring surface tension of solutions has been criticized or rejected by some group of investigators, the general picture must be substantially correct. In our opinion the most certain method of investigation, although not the most precise, is the P.L.A.W.M. trough[15]. This is the only method in which the surface is plane, completely undisturbed during measurement, wholly unaffected by contact angle or by adsorption on the instrument, and in which the surface is demonstrably swept clean, the absence or degree of contamination being directly measured at every desired stage. The P.L.A.W.M. trough again demonstrates the reality of type III curves, as well of course as of types I and II.

Milner[7] thought that the contradiction between the theory and experiments on surface tension and adsorption might be due to the surface tension being irreversible and therefore inadmissible, but this conjecture is not substantiated by

test in the Langmuir trough[14], and in any case would not necessarily alter the value from that of the equilibrium surface tension.

Another frequently suggested reason for rejecting the applicability of Gibbs's theorem is that many of these systems are said to contain colloids. However, this is a very serious attack upon the validity of thermodynamics in general, because all these solutions have physical chemical properties that are defined solely by temperature and composition. They have a definite lowering of freezing point, a definite vapour pressure at each temperature, etc.; and they enter into true reversible equilibria with other phases. The applicability of thermodynamics is therefore inescapable. The solutions possess or exhibit definite values of chemical potential for each of the components. Furthermore, the chemical potential of each component must rise with increase in concentration of that component, and at every point the chemical potential of each component in the surface must be identical with that in the solution. Hence neither can pass through maxima or minima without the solution itself being unstable. This is a point sometimes overlooked in attempting to account for adsorbed sodium oleate in excessive dilution.

In this connexion it is of importance always to remember that the chemical potentials or activities of the components are those which are measured in the bulk of the solution itself, and have nothing to do with surface phenomena, except that for equilibrium they must be uniform throughout the system. As is vividly portrayed by G. N. Lewis, when the surface area is increased the Gibbs adsorption is the amount of substance which has to be added to the remainder of the system to restore the activity to its original value at any point within the bulk of the solution.

Thus in discussing the chemical potential of a solution no reference need be made to the surface; whereas in discussing the structure and composition of the surface there is at once a need for comparing it with the bulk of the solution and explaining by what mechanisms the surface tension is regulated and the chemica potentials are adjusted to equal that of the solution. Similarly in reversible colloidal systems, the attention may be directed either to the definite properties of the solution as a whole, or to the composition, structure, and electrification of a particle, the activities of whose components are necessarily equal to that prevailing in the solution as a whole.

Another reason for not applying the Gibbs theorem is the assumption that in certain cases the system contains more than two components. Gibbs's general equation of course does supply a term for each and every component actually present. A case where it is not legitimate to regard the solution as a two-component system is where hydrolysis occurs, as in an oleate where one of the products is partially rejected from the surface, and is distributed throughout the solution However, if both products of hydrolysis are present in equal amounts in or near the surface, then the solution is still in the Gibbsian sense a two-component system.

Similarly, if one ion is preferentially sorbed, but the other remains in the neighbourhood of the surface, as part of either a Helmholtz or a diffuse double layer, the system still contains only two components. In Gibbs's derivation, it is expressly stipulated that the surface includes the whole zone, however deep,

between the completely homogeneous portions of the vapour and liquid phases. Thus dissociation does not affect the theorem as long as the system remains electrically neutral.

It must now be pointed out that although some of the systems quoted in the appendix are undoubtedly affected by hydrolysis, this is not the case for such substances as the free acids, the free bases, and a number of non-electrolytes giving type III curves. The non-electrolytes tend to give enormous lowering of surface tension in extreme dilution to a constant value which is maintained over all higher concentrations, corresponding according to the theorem to no adsorption at all, except just in extreme dilution where the solute is present in inappreciable quantity! It is easy to demonstrate that adsorption often actually occurs where none is predicted (Ts-Ming Woo, unpublished experiments) by comparing the concentration of collapsed foam with that of the solution with which it was in contact. Likewise, the microtome method shows positive adsorption of sodium butyl benzene sulphonate in the region where the surface tension is constant (L. A. Wood, unpublished experiments).

Another suggestion has been that the solute is in a different state of association or ionization in the surface from that in the bulk of the solution, and therefore it behaves as more than one distinct substance. This is contrary to the whole treatment of thermodynamics, as for example, in discussing the solubility of salts in liquids, or such phenomena as the distribution of acetic acid between water and the associating solvent benzene. It would also deny the applicability of the phase rule to the system ice, water, water-vapour, a serious restriction of the validity of chemical thermodynamics. So long as the various forms of any substance are rapidly and reversibly interconvertible and are in true equilibrium with one another, they constitute only one component.

A similar far reaching invalidation of thermodynamics would follow from the recent suggestion that the Gibbs theorem applies only to ideal solutions of a single component. The older suggestion that the Gibbs theorem applies solely to a fraction of the depth of the surface is likewise clearly at variance with the explicit basis of the Gibbs formulation.

§ 5. GIBBS'S THEOREM AS A LIMITING LAW

The Gibbs theorem would appear to be like the ideal gas law in that the latter omits factors such as those represented by a and b of van der Waals's equation. For many type I curves the theorem may represent a close approximation to the facts.

However, the fact that the largest observed lowerings of surface tension of water occur just where the Gibbs adsorption is zero proves that a most important factor or factors has been omitted. Such may be orientation, submerged double layer, or free electrical charge.

The enormous negative adsorptions demanded by the Gibbs theorem, and the only thermodynamically admissible explanation, agree in showing that the surface

of a solution of a soluble substance is very deep. Here the surface tension conditions and is conditioned by molecules and ions that are far below the outside of the surface. The great depth of surfaces of solutions of soluble substances contrasts strongly with the familiar conceptions of the simpler systems of insoluble monolayers, although it becomes apparent that the latter must be somewhat modified in view of the facts here adduced, and that the application of the Gibbs theorem to insoluble oil films on water is only an approximation. Likewise the use of the Gibbs equations to provide the basis for an analogy between surfaces of solutions and two-dimensional monolayers is precarious and may in many cases be misleading.

For example, a monolayer of oleic acid lowers the surface tension of water only to 40 dynes whereas that of pure oleic acid is 32·5 dynes with the same exterior. Sodium oleate lowers the surface tension of water to as little as 23 dynes, yet with the same exterior as the other two liquids. Hence surface tension is not determined solely by the exterior layer of molecules.

In the case of soap and probably in every surface of water or other dissociating solvent there must be a submerged double layer extending in many cases hundreds or thousands of Ångström units deep and yet reducing the surface tension through repulsion of charges of like sign and attraction of opposites. This accounts for the far lower surface tension of highly dissociated soap solutions as compared with the weak fatty acids. Yet it is well known that even insoluble fatty acid films exhibit some dissociation[16, 17] and therefore to that extent develop a submerged double layer which must account for part of the lowering observed with insoluble films. The lowering due to double layer is evidently about half the total lowering for soap solutions and only a minor effect with insoluble fatty acids. This factor must exist even at the surface of pure water where electrokinetic movement and initial repulsion of bubbles and of droplets is observed[18].

This new conception of the limitations of Gibbs's theorem in its usual form was made more familiar in a previous communication[15] for a two-component system such as an aqueous solution of a non-hydrolysable hydrogen soap. Here the outside of the surface begins with the terminal methyl groups of the paraffin chain, then the rest of the negative ion, or molecule, then free hydrate water; then the hydrogen ions, some close to the anions and others interspersed between water and soap molecules and ions in the familiar diffuse Gouy zone. The latter has been formulated by Gouy and others; and equivalent formulae were derived in the Donnan equilibrium, and in the Debye-Hückel theory of interionic attraction. Probably the total surface is some thousands of Ångström units deep before the completely normal state of the bulk is reached. This has been directly proved for particular cases by x-ray and electron diffraction by Trillat[19], Clark[20], Tausz[21] and others, in accordance with observations on lubrication, adhesion and thixotropy.

Another reason for regarding a surface as being more than one molecule deep was found by Goard and Rideal[22] in that sodium chloride raised the surface tension of phenol solutions in a manner similar to that of pure water.

The two most powerful influences on surface tension are: first, the exposed terminal groups or molecules as emphasized by Langmuir, and secondly, the sub-

merged double layer[15]. They are partly independent and may even operate to oppose as well as interfere with each other.

Finally, the submerged double layer may be adduced to account for the minimum in type III curves of electrolytes, for in increasing concentrations the dissociation will diminish and allow surface tension to rise.

§ 6. THE ADDITIONAL TERMS AFFECTING SURFACE TENSION[28]

It is agreed that long range repulsive forces will arise from the mutual repulsion of similarly oriented induced or permanent dipoles upon the surface. This must occur in one-component systems even in the absence of Gibbs's adsorption.

Such dipoles will preferentially attract, in the layers underneath, ions of one sign rather than the other; forming a submerged double layer with its powerful electrocapillary effect in reducing surface tension. Langmuir[23], Magnus and others have discussed formal expressions for the summation of the primary effect of mutual orientation of undissociated dipoles or quadrupoles. If it includes also the shorter range cohesive forces, the term may be positive or negative.

Similarly a second term should be added for the electrocapillary effect of the submerged double layer present in aqueous systems. This also has been considered and formulated in connexion with the subject of electrocapillarity, although mainly for the special case of the liquid metallic conductor, mercury. A recent discussion with many references is given by Gatty and McKay[4].

Both of these phenomena must, like the surface tension, be a function of the chemical potentials of the bulk of the solution.

Finally, if, as seems frequently to happen, the system is actually electrified, a third independent term of the form used elsewhere by Gibbs must be added: $e\,dV$, where e is the quantity of electricity and V the potential. This term would not be a function of the chemical potential of the solution as normally measured. Each of the factors enumerated in this extended form of equation (508) can influence all of the others; their interrelation is determined by varying but two at a time or by calculating from other evidence the magnitude of any one or more of them.

Gibbs's equation (497), as Gibbs himself points out, is like his equation (12) or indeed his equation (11) in which the quantities of matter are not regarded as variables. These are

$$d\epsilon = t\,d\eta - p\,dv, \qquad \text{.....(11)}$$

$$d\epsilon = t\,d\eta - p\,dv + \mu_1 dm_1 + \ldots + \mu_n dm_n. \qquad \text{.....(12)}$$

All three equations (11), (12) and (497) are expressions of the first law of thermodynamics or the conservation of energy. What they exclude is as important as what they include because there is an explicit term for every kind of energy that is taken into account, and it is a rigorous requirement that no other energy change is taking place.

The suggestion is then that further terms should be inserted in equation (497) and in the final equation (508) to account for effects of orientation, of submerged

double layer, and of free electrical charge, if any. Just as Gibbs had explicitly in view the change in surface tension due to change of amount of material in the surface, the surface tension may be altered by change of orientation of molecules within the surface even when no change in composition or concentration has occurred; that is, in the absence of "adsorption". This might well be discussed for a one-component system, for example, following Gibbs's suggestion of compensation at constant mass and temperature by a suitable change in curvature (p. 230). Equally well may the surface tension be altered by arrangement of existing ion pairs to form a double layer or a diffuse double layer, even in the absence of adsorption.

§ 7. APPLICATION OF THE FILM BALANCE TO ORDINARY SOLUTIONS

Contrary to the expectation of McBain and Wilson[14] and of Doss[24], the film balance definitely does not measure surface tensions of solutions. It has been shown by McBain and Perry[25], that for some solutions there is no effect upon the moveable float when the surface on one side of the float is expanded or compressed. Indeed, this is frequently the case for solutions of any other substances shortly after their surfaces have been swept. Likewise, Ford and Wilson[26] found that in general oleic acid spreads as rapidly upon freshly swept surfaces as upon pure water. In all these cases the surface tension is already lowered far below that of water and yet this escapes detection by these methods of examination.

However, after the surface has been allowed to age or mature, which may require an hour or more, there are three classes of behaviour observed in the film balance. Either there is no effect on the float upon compression of the surface or there is a film pressure evidenced which quickly vanishes or in many cases, even of comparatively simple soluble substances such as caproic acid or beta-phenylpropionic acid, a hitherto unsuspected phenomenon appears. In this last case there is no effect upon the float until the surface has been compressed to a definite fraction dependent upon the concentration, at which point a temporary pressure suddenly appears and then vanishes comparatively slowly. Upon somewhat greater compression a greater pressure is produced. This does not vanish but falls to a final characteristic value which is independent of the concentration and dependent only on the substance being examined. The same final characteristic value is obtained when a droplet or small crystal of the substance is placed upon pure water on one side of the film balance. In other words, upon compressing or collecting a matured surface a compacted semi-insoluble or semi-denatured film is found which requires a definite pressure to be overcome before it will dissolve. For hydrocinnamic acid the value is always 0·42 dyne/cm. This represents a definite false equilibrium for a compressed surface. As stated above, the characteristic value of many substances is zero.

These novel facts as well as the many recent studies (for example Adam and Shute[27]) which show the long time, often amounting to hours or weeks, required for the surface tension of the solution to assume a definite value are of great signifi-

cance both in applying and interpreting the Gibbs theorem. They accord well with the foregoing deduction that the surfaces of solutions are deep. The semi-denatured film that forms and lies upon the surface is however more probably monomolecular until compressed.

§ 8. SUMMARY

Abundant examples of type III surface tension curves have been obtained by many authors by every experimental method (see appendix). In these, the static or true equilibrium surface tension is very greatly lowered in extreme dilution, thereafter remaining constant or passing through one or more maxima or minima, although the dynamic or immediate surface tension is several times greater, nearly that of water. Application of the classical Gibbs theorem in its exact form to these cases yields a series of obviously impossible results, no matter in what form they are calculated. The different reasons adduced by many authors for avoiding such applications are seen to be gravely at variance with the foundations of chemical thermodynamics.

It follows from the Gibbs equation that the surface of solutions of soluble substances is very deep as compared with molecular dimensions, in contrast to the familiar conception of insoluble films on water. This is to some extent substantiated by the long time required to obtain equilibrium surface tensions for solutions and is also in accord with the novel phenomena recently discovered by applying the film balance and by measuring the rate of spread of oleic acid on solutions.

The classical Gibbs adsorption theorem appears to be a limiting law applying to cases where mutual repulsion or oriented dipoles upon the surface and the effects of submerged double layer and of electrification do not also condition the surface tension.

REFERENCES

(1) GUGGENHEIM, E. A. and ADAM, N. K. *Proc. Roy. Soc.* A, **139**, 218 (1933). (Alcohol in water gives a type I curve.)
GUGGENHEIM, E. A. *Modern Thermodynamics by the Method of Willard Gibbs* (Methuen and Co. Ltd., London, 1933).
Cf. LEWIS and RANDALL. *Thermodynamics*, p. 250 (McGraw-Hill Book Company, Inc. N.Y. 1923).
(2) KOENIG, F. O. *J. Phys. Chem.* **38**, 11 (1934).
(3) DEFAY, R. *Étude Thermodynamique de la Tension Superficielle*, p. 193 (Gauthier-Villars and Company, Paris, 1934).
(4) GATTY, O. and MCKAY, H. A. C. *Phil. Mag.* (7), **22**, 402 (1937).
CRAXFORD, S. R., GATTY, O. and THEORELL, T. *Phil. Mag.* (7), **25**, 104 (1938).
(5) GIBBS, J. WILLARD. *Collected Works*, **1**, 228 and 230 (Longmans Green and Co. New York, London, Toronto, 1928).
(6) VAN RYSSELBERGHE, P. *J. Phys. Chem.* **39**, 403 (1935).
(7) MILNER, S. R. *Phil. Mag.* (6) **13**, 98 (1907).
(8) BOND, W. N. and PULS, H. O. *Phil. Mag.* (7) **24**, 864 (1937).
(9) FOULK, C. W. *Kolloidzschr.* **60**, 115 (1932).
(10) MCBAIN, J. W. and HUMPHREYS, C. W. *J. Phys. Chem.* **39**, 300 (1935).
MCBAIN, J. W. and SWAIN, R. C. *Proc. Roy. Soc.* A, **154**, 608 (1936).
(11) JONES, G. and RAY, W. A. *J. Amer. Chem. Soc.* **59**, 195 (1937).
See also EVANS, A. W. *Trans. Faraday Soc.* **33**, 794 (1937).

(12) HARKINS, W. D., DAVIES, E. C. H. and CLARK, G. L. *J. Amer. Chem. Soc.* **39**, 549 (1917).
(13) EDGERTON, H. E., HAUSER, E. A. and TUCKER, W. B. *J. Phys. Chem.* **41**, 1017 (1937).
(14) MCBAIN, J. W. and WILSON, D. A. *J. Amer. Chem. Soc.* **58**, 379 (1936).
(15) MCBAIN, J. W., FORD, T. F. and WILSON, D. A. *Kolloidzschr.* **78**, 1 (1937).
(16) MCBAIN, J. W. and PEAKER, C. R. *Proc. Roy. Soc.* A, **109**, 28 (1925).
(17) MCBAIN, J. W. and FOSTER, J. F. *J. Phys. Chem.* **39**, 331 (1935).
(18) GIBBY, C. W. and ADDISON, C. C. *J. Chem. Soc.* 119, 1306 (1936).
(19) TRILLAT, J. J. *Rev. Sci.* **64**, 552 (1926); *Ann. Phys., Lpz.*, **10**, 5 (1926); *C.R. Acad. Sci., Paris*, **182**, 843 (1926); **187**, 168 (1928); *Mech. Engr*, **50**, 471 (1928); *Hall a. cuirs*, Suppl. Tech. 1927, 137; *Metallwirtschaft*, **7**, 101 (1928).
(20) CLARK, G. L., LINCOLN, B. H. and STERRETT, R. R. *Proc. Amer. Petrol. Inst.* Sect. III, **16**, 81 (1935).
(21) TAUSZ, J. *et al. Z. angew. Chem.* **43**, 570 (1930); *Erdöl u. Teer*, **9**, 331 (1933).
(22) GOARD, A. K. and RIDEAL, E. K. *J. Chem. Soc.* **127**, 1675 (1925).
(23) LANGMUIR, I. *J. Amer. Chem. Soc.* **54**, 2798 (1932); *Gen. Elect. Rev.* **38**, 402 (1935).
(24) DOSS, K. S. G. *Curr. Sci.* **4**, 405 (1935); **5**, 645 (1937); *Proc. Indian Acad. Sci., Bangalore*, **4**, 11, 97 (1936); *Kolloidzschr.* **84**, 138 (1938).
(25) MCBAIN, J. W. and PERRY, L. H. *Industr. Engng. Chem.* **31** (Jan. 1939).
(26) FORD, T. F. and WILSON, D. A. *J. Phys. Chem.* **42**, 1051 (1938).
(27) ADAM, N. K. and SHUTE, H. L. *Trans. Faraday Soc.* **34**, 758 (1938).
(28) Cf. the recent discussions, from very different points of view, of various electrical effects affecting surface tension and its measurement, and adsorption, by M. Dole, *J. Amer. Chem. Soc.* **60**, 904 (1938); J. J. Bikerman, *Trans. Faraday Soc.* **34**, 1268 (1938); and I. Langmuir, *Science*, **88**, 430 (1938); contrasting with others who have expressed the opinion that the Jones-Ray effect is merely due to accidental contamination.

Appendix. References to papers dealing with type III curves, for aqueous solutions whose surface tensions are greatly lowered in very dilute solution, thereafter passing through a minimum or remaining constant.

(1) BELTON. *Trans. Faraday Soc.* **31**, 1421 (1935); maximum bubble pressure method; ethyl alcohol-water-salt mixtures, using LiCl, NH_4Cl, and KCNS.
(2) BHATNAGAR, PRASAD and BAHL. *Quart. J. Indian Chem. Soc.* **2**, 11 (1925); drop number, sodium oleate.
(3) CUPPLES. *Industr. Engng. Chem.* **27**, 1219 (1935); du Noüy, sodium hydroxideoleic acid mixtures.
(4) EDGERTON, HAUSER and TUCKER. *J. Phys. Chem.* **41**, 1017 (1937); drop weight, sodium oleate.
(5) EVERSOLE and DEDRICK. *J. Phys. Chem.* **37**, 1205 (1933); drop volume; effect of sodium formate and acetate on interfacial tension of water-mineral oil and water-toluene systems; also perchloric acid.
(6) EVANS. *Trans. Faraday Soc.* **33**, 794 (1937); drop volume; effect of NaOH and LiCl on interfacial tension of *n*-hexane and water.
(7) GIBBY and ADDISON. *J. Chem. Soc.* **126** (1936); drop number, Congo red.
(8) HARKINS, DAVIES and CLARK. *J. Amer. Chem. Soc.* **39**, 586 (1917); drop weight, sodium oleate; also at benzene-water interface, effect of sodium hydroxide and barium hydroxide.
(9) JOHLIN. *J. Phys Chem.* **19**, 1129 (1925); capillary rise, sodium oleate.
(10) JONES and RAY. *J. Amer. Chem. Soc.* **59**, 195 (1937); capillary rise, potassium chloride, potassium sulphate and cesium nitrate.
(11) LASCARY. *Kolloidzschr.* **34**, 73 (1924); drop number, sodium caproate, caprylate, laurate and oleate.
(12) LEHNER and BUELL. *Industr. Engng. Chem.* **8**, 701 (1916); drop number, sodium oleate.

(13) LOTTERMOSER and GIESE. *Kolloidzschr.* **73**, 279 (1935); modified torsion balance; potassium and lithium caproate, caprylate, caprate, laurate, myristate, palmitate, stearate, oleate, and abietate.

(14) LOTTERMOSER and STOLL. *Kolloidzschr.* **63**, 57 (1933); torsion balance; sodium dodecyl, tetradecyl, hexadecyl, octodecyl sulphate; silver, calcium, magnesium, iron and copper dodecyl sulphates; silver and calcium tetradecyl sulphates.

(15) LOTTERMOSER and WINTER. *Kolloidzschr.* **66**, 286 (1934); du Noüy, sodium laurate, myristate, palmitate, stearate, oleate; potassium palmitate, oleate, and stearate; lithium palmitate and stearate; triethanol-palmitate, stearate and oleate; tragacanth, gum arabic, saponin and gelatine.

(16) MILNER. *Phil. Mag.* (6) **13**, 101 (1907); capillary rise, sodium oleate.

(17) NARAYAN and SUBRACHMANYAM. *Phil. Mag.* (6) **43**, 663 (1922); maximum bubble pressure and capillary rise; sodium oleate.

(18) NEVILLE and JEANSON. *J. Phys. Chem.* **37**, 1002 (1933); torsion balance; sodium isopropylbenzene, *p*-methyl-isopropylbenzene and butylbenzene sulphonates; and gardinol.

(19) NICKERSON. *J. Phys. Chem.* **40**, 285 (1936); capillary rise, acid sodium oleate.

(20) NISHIZAWA, WINOKUTI and KIKUTI. *Tech. Rep. Tohoku Imp. Univ.* **10**, no. 1, 107 (1931); drop number; sodium ricinoleosulphate, sodium ricinoleate.

(21) NUTTALL. *J. Soc. Chem. Ind.* **39**, 67 T (1920); drop number; sodium oleate.

(22) *Physical and Chemical Properties of Rubber* (published 1935 by Research Association of British Rubber Manufacturers, Croydon, England, p. 7); torsion balance; rubber latex.

(23) POWNEY and ADDISON. *Trans. Faraday Soc.* **33**, 1246, 1254 (1937); du Noüy; sodium dodecyl, tetradecyl, hexadecyl and octodecyl sulphates; and effect of added sodium and calcium chlorides.

(24) RAYLEIGH. *Phil. Mag.* (5) **30**, 386 (1890); *Proc. Roy. Soc.* A, **47**, 281 (1890); capillary rise, sodium oleate.

(25) REED and TARTAR. *J. Amer. Chem. Soc.* **58**, 322 (1936); capillary rise, sodium *n*-octyl and *n*-decyl sulphonates.

(26) SALAZAR. *C.R. Acad. Sci., Paris*, **201**, 1120 (1935); capillary rise; malachite green.

(27) WALKER. *J. Chem. Soc.* **119**, 1521 (1921); drop number; sodium laurate, myristate and palmitate.

(28) WARK. *J. Phys. Chem.* **40**, 666 (1936); maximum bubble pressure; trimethyl cetyl ammonium bromide.

(29) WINOKUTI, IGARASI and YAGI. *Tech. Rep. Tohoku Imp. Univ.* **10**, no. 3, 73 (1931); drop number; sodium ricinoleosulphate, ricinoleate, nd oleate.

(30) WOO, unpublished results, du Noüy. With one or more minima: potassium oleate, an Igepon T, a purified Igepon T. With horizontal portions: technical sodium oleyl sulphate, technical sodium octyl sulphate, technical sodium stenyl sulphate, nonyl glucoside, polyhydroxyethyl ether, technical sodium alkyl sulphate, ammonium alkyl sulphate, an Igepon A, polyglycerol ester, mono-sodium sulphonate of dioctyl succinate, oleoglyceryl sulphate, diglycol laurate, triethanolamine, synthetic non-electrolytic polymer.

(31) WOOD, unpublished results; du Noüy. With one or more minima: sodium butyl benzene sulphonate, butyl benzene sulphonic acid.

RECENT THEORIES OF THE LIQUID STATE

By N. F. MOTT and R. W. GURNEY
H. H. Wills Physical Laboratory, University of Bristol

§ 1. INTRODUCTION

During the past few years the use of x-ray diffraction has enabled a considerable amount of information to be obtained about the structure of liquids. In particular, it has shown that the molecules of a liquid have a certain degree of order, as in a solid, and are not distributed entirely at random, as in gas. It has therefore become fashionable to think of a liquid rather as a disordered or broken-up solid than as a highly compressed gas. The adoption of this point of view has led to a number of new attempts to deduce the equation of state, freezing and boiling points and latent heats of liquids in terms of the properties of the atoms of which they are made up. It is the chief purpose of this article to give an account of these attempts. In all these theories the aim must be to calculate the free energy F as a function of the volume and temperature. The equation of state is then given by

$$p=\left(\frac{\partial F}{\partial v}\right)_T,$$

and the melting point by the condition that the free energy of the liquid should be equal to that of the solid. In practice it is usually more convenient to calculate the "partition function" f, defined by

$$f=e^{-F/kT}.$$

We shall also give a short account of some recent attempts to understand flow in liquids, and therefore to obtain formulae for the viscosity and diffusion coefficients. Throughout this article we shall confine ourselves to monatomic liquids, and liquids with spherical molecules. We include in a final section a discussion of some theories about liquid helium.

§ 2. THE PARTITION FUNCTION OF A SIMPLE LIQUID, CONSIDERED AS COMPRESSED GAS

Statistical mechanics has developed to the point where the theory yields fairly accurate partition functions for solids and vapours. This is not yet possible for liquids; all recent attempts to find approximate partition functions treat a liquid either as a highly condensed gas, or more often as a very imperfect solid. In this section we shall briefly describe the method of obtaining the equation of state of an imperfect gas and the attempts that have been made to extend these methods to the densities characteristic of the liquid phase.

Throughout this article we shall consider a substance made up of N molecules of mass m, and shall suppose for simplicity that the temperature is so low that the internal degrees of freedom of each molecule may be disregarded. Then, whether the substance is in the solid, liquid or gaseous phase, the partition function is given by

$$f=\frac{1}{N!}\left[\frac{(2\pi mkT)^{\frac{3}{2}}}{h^3}\right]^N \int e^{-W(q_1,q_2,\ldots)/kT}\,dq_1 dq_2\ldots, \qquad \ldots\ldots(1)$$

where $W(q_1, q_2, \ldots)$ is the potential energy of the substance when the molecules are in positions defined by the co-ordinates $q_1, q_2, \ldots$. The factor $N!$ expresses the fact that the particles are indistinguishable, and obey either Einstein-Bose or Fermi-Dirac statistics; though in either case we assume that we are above the degeneracy temperature, so that classical formulae are applicable.

Assuming central forces between the molecules, W can be expressed as a double sum

$$W=\tfrac{1}{2}\sum_i\sum_k w(r_{ik}),$$

where $w(r)$ is the potential energy of a pair of molecules at a distance r from one another, and the summation is over all pairs of molecules. In the usual treatment of imperfect gases, it is assumed that the configurations in which groups of more than two molecules are in one another's field at any moment are rare, and make a negligible contribution to (1); thus the integral can be replaced by the product of a number of integrals of the type

$$\int e^{-w(r)/kT}\,4\pi r^2\,dr, \qquad \ldots\ldots(2)$$

each integration being over all space, since by our assumption all of the space in which $w(r)$ is finite is available to the two molecules. The evaluation of the partition function and the deduction of the equation of state leads to van der Waals's equation in the form

$$pv=RT\left(1+\frac{b}{v}\right),$$

the constant b, known as the first Virial coefficient, being given in terms of the integral (2). Attempts[1] have been made to obtain formulae which take into account ternary and higher encounters, and thus to obtain a more accurate equation of state in the form

$$pv=RT\left(1+\frac{b}{v}+\frac{c}{v^2}+\ldots\right).$$

Doubts[7] have, however, been expressed as to the convergence of series obtained in this way for volumes in the neighbourhood of the melting point. A physical reason may be given for the insufficiency of such a method of approximation. In a gas it may perhaps be possible, starting from any one molecule, to say that only one, two or three other molecules are within its field of force, or within the field of force of molecules which are within its field of force and so on. If we can say that negligibly few groups of more than n molecules exist which are interacting with one

another, then n terms of the series ought to be sufficient. But this is clearly not possible for a liquid, which forms in effect one single cluster. In a liquid, if we start from any one molecule we can reach *all* the others without passing any gaps so great that the intermolecular forces cannot act across them.

§ 3. THE EINSTEIN MODEL FOR A LIQUID

In the work carried out during the last few years by Eyring[2,3,4,5,6], Lennard-Jones[7,8], Bernal[9] and others, it has been usual to emphasize the similarity between the liquid and solid phases, rather than that between the liquid and the vapour. All these authors emphasize the fact that a molecule within a liquid is *all the time* within the field of force of about as many other molecules as in the solid phase; it does not, as in a gas, spend most of its life in comparative freedom and interact strongly with other molecules only occasionally. As Lennard-Jones has expressed it, an atom is regarded as imprisoned in a cell for most of its time by its immediate neighbours, just as it is in a solid, and though migration or diffusion will in practice take place from one cell to another, in calculating the partition function and hence the equilibrium properties of the liquid this possibility is neglected as being an infrequent event compared with the duration of time spent within each cell.

The partition function, for liquid, solid or gas, is given by formula (1). For a solid there are various ways in which approximate evaluations of the integral may be made, and in order to understand the analogous calculations for liquids it will be as well to review one of these. In the model used by Einstein in his theory of specific heats, we divide up the energy of the solid into the energy $\phi_0(v)$ of the atoms at rest in their mean positions with atomic volume v, and the potential energy due to their displacements from their mean positions. In the simple model the atoms are supposed to be able to vibrate independently of one another with a frequency ν (itself a function of v), $h\nu/k$ being identified with the characteristic temperature of the solid. The term in the potential energy due to the displacements is thus

$$V(r_1, r_2, \ldots) = 2\pi^2 m\nu^2 \sum_k (x_k^2 + y_k^2 + z_k^2), \qquad \ldots\ldots(3)$$

where $r_k = (x_k, y_k, z_k)$ is the displacement of the kth atom.

The integral in (1) then becomes

$$N! \left[\int e^{-V(r)/kT} 4\pi r^2 \, dr \right]^N e^{-\phi_0/kT}. \qquad \ldots\ldots(4)$$

The factor $N!$ occurs because there are $N!$ ways in which the N atoms of the solid can be distributed among the N positions of equilibrium. (4) can be evaluated to give

$$N! \left[\left(\frac{kT}{2\pi m\nu^2} \right)^{\frac{3}{2}} \right]^N e^{-\phi_0/kT},$$

so that the partition function becomes

$$\left(\frac{kT}{h\nu} \right)^{3N} e^{-\phi_0/kT}.$$

Since the free energy is given by

$$F = U - TS = -kT \log f,$$

we have

$$F = \phi_0 + 3NkT \log \frac{h\nu}{kT},$$

which is the usual expression for the free energy of a simple Einstein solid[26]. The second term accounts for the thermal expansion.

If we try to treat a liquid in the same way, we are faced by the following difficulties:

(i) Owing to the different arrangement of the atoms in a liquid, neither ν nor ϕ_0 will be exactly the same as in the solid, and except near the melting point the motion of the atoms will not be simple harmonic.

(ii) There will be a certain "entropy of disorder" which must be included in the free energy.

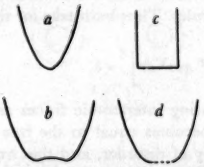


Figure 1. Potential energy of atom in field of other atoms. (*a*) Solid. (*b*) Liquid (Lennard-Jones). (*c*), (*d*) Liquid (Eyring).

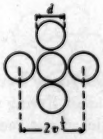


Figure 2. Free volume in a liquid, according to Eyring.

As regards (i), the calculations which seem most satisfactory are those of Lennard-Jones[7, 8]. As already explained, he assumes each atom to be moving independently of the others in a cell surrounded by the other atoms. For ϕ_0 he takes effectively the energy of the solid, namely the mutual potential energy of the assembly when each atom is at the centre of its cell. Corresponding to (4), he takes, as one would in a solid using Einstein's model, the potential of an atom when all the neighbouring atoms occupy their mean positions; but he calculates this for inert gases using the law of force obtained from other data. Of course for temperatures far from the melting point this does not give a simple harmonic form at all, but a potential such as illustrated in figure 1 (*b*).

Eyring also separates the energy of the liquid into a term $\phi_0(v)$ and a term depending on the displacements of the atoms. His treatment of the latter is cruder; the atoms are treated as hard spheres of diameter d. If in the liquid at any pressure the atomic volume is v, the atoms are represented as moving in a "free volume" v_f given by

$$v_f = 8\,(v^{\frac{1}{3}} - d)^3.$$

This is the volume available for a molecule if the surrounding molecules are considered as being at rest in their mean positions as shown in figure 2.

This may be a good approximation far from the melting point, but near the melting point, where the heat capacity per gramme molecule is near to $3R$, it is probable that the motion of the atoms is more nearly simple harmonic. In view of this, in a discussion of liquid mercury, Eyring has considered each atom to move in a field of the type illustrated in figure 1(*d*), the curved parts moving apart as the atomic volume is increased. In another paper he considers d to be a function of temperature. Eyring's work in general attempts to correlate the different properties of liquids in terms of a simple model, rather than to deduce them from the properties of atoms.

§ 4. THE ENTROPY OF A LIQUID

It is to be emphasized, then, that apart from a term giving the entropy of disorder, both Lennard-Jones and Eyring take the partition function of a liquid to be exactly like that of a solid when the Einstein model is used, except that they do not assume harmonic vibrations for the molecules. Each molecule is regarded as imprisoned in a cell by the surrounding molecules. Thus both take for the partition function

$$f=\left[\frac{(2\pi mkT)^{\frac{3}{2}}}{h^3}\int e^{-V(r)/kT}\,4\pi r^2\,dr\right]^N\times\delta, \qquad \text{.....(5)}$$

and Lennard-Jones works out the integral using interatomic forces as explained above, while in Eyring's work the integral becomes equal to the free volume v_f per molecule. δ is a factor giving the entropy of disorder, and this we shall now discuss.

Both authors assume that $\delta=N^N/N!$. .....(6)

Since by Stirling's theorem $\log\dfrac{N^N}{N!}\sim N$,

and since the entropy S is equal to $-U/T+k\log F$, it follows that the factor (6) corresponds to an entropy increase of Nk, or R per gramme molecule. It may be remarked here that the entropy of melting (L/T_M, where L is the latent heat and T_M the melting point) is of this order of magnitude.

The factor $N^N/N!$ may be understood in the following way. Consider a volume V which is divided into N equal cells, each of volume V/N. Suppose now we have N molecules, and put one in each cell. The partition function is then, by (1),

$$\left[\frac{(2\pi mkT)^{\frac{3}{2}}}{h^3}\right]^N\left(\frac{V}{N}\right)^N. \qquad \text{.....(7)}$$

The factor $N!$ has disappeared, as before, because there are $N!$ ways in which the molecules can be distributed among the cells.

The partition function (7) is of the "solid" type; if we take an Einstein solid, but, instead of letting each molecule vibrate with simple harmonic motion, give it a free volume v_f to move in, the partition function is

$$\left[\frac{(2\pi mkT)^{\frac{3}{2}}}{h^3}\right]^N v_f{}^N. \qquad \text{.....(8)}$$

Now let us remove the partitions between the cells, so that we have in fact a perfect gas of N molecules in a volume V. Then, by (1), the partition function is

$$\left[\frac{(2\pi mkT)^{\frac{1}{2}}}{h^3}\right]^N \frac{V^N}{N!}, \qquad \text{.....(9)}$$

which is the ordinary partition function of a perfect gas. This is larger by the required factor $N^N/N!$ than (7). Thus by removing the partitions between the cells, we have increased the entropy by Nk. This entropy arises from the fact that, now, any number of molecules may be placed in each cell. The factor $N^N/N!$ may be obtained at once from this consideration, since the number of ways that N molecules can be distributed among N cells with no restriction on the number in each cell is N^N, while if only one can be placed in each cell the number is $N!$.

Consider now an imperfect gas of N molecules moving in a volume V; for simplicity let them be considered as hard spheres. Suppose that in working out the integral (1), we take all the atoms except one to be at rest in arbitrary positions and

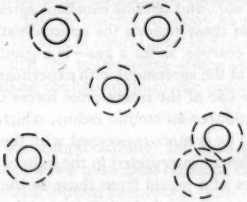


Figure 3.

Figure 4.

find the volume V_f available to the one atom considered. This is the volume outside the broken spheres in figure 3. This volume will actually depend on how the other atoms are distributed; e.g. if two atoms are in contact, the free volume V_f increases. Let then $\overline{V}_f$ be a mean value of V_f. Then it is a reasonable assumption that under certain conditions an approximation to the partition function may be

$$\frac{1}{N!}\left[\frac{(2\pi mkT)^{\frac{1}{2}}}{h^3}\right]^N (V_f)^N. \qquad \text{.....(10)}$$

Defining the free volume per molecule v_f as $\overline{V}_f/N$, this becomes

$$\left[\frac{(2\pi mkT)^{\frac{1}{2}}}{h^3}\right]^N \frac{N^N}{N!}\, v_f{}^N. \qquad \text{.....(11)}$$

Comparing this with the "solid" partition function (8), we see how the factor $N^N/N!$ has been introduced.

The condition, however, that (10) shall be a good approximation to the partition function is clearly that V_f for one atom shall *not* depend much on the positions of the other atoms. But when, as is done by Eyring, the factor $N^N/N!$ is combined with an Einstein model of a liquid, this is by no means the case. The whole assumption made by him is that at any one moment atom A, figure 4, is moving in a free

volume v_f, which for the moment is defined by the positions of the other atoms, and *if it is to move into the other free volumes available, the other atoms must move into new positions*. It seems to the present authors that there will be few, if any, configurations of the whole system for which the free volume for any one atom is as great as Nv_f. Thus if v_f is properly defined (as in Lennard-Jones's work) for the Einstein liquid, we believe that the factor $N^N/N!$ is too large, and gives a worse approximation the nearer we are to the melting point.

A further criticism of the use of this factor is that it does not allow any place for the concept of a varying degree of order, and rules out from the beginning the possibility of any continuous transition from solid to liquid phase.

Before reviewing those theories in which a changing degree of order is taken into account, we shall discuss the agreement with experiment obtained by Lennard-Jones and by Eyring in the work described above. Lennard-Jones uses his partition function to calculate the properties of condensed gases far from the freezing point; for instance, he calculates the critical temperature, Boyle temperature and boiling point of the rare gases, oxygen, nitrogen, etc., and obtains excellent agreement with experiment. It appears, then, that in just these regions the approximation used—namely, energy calculated as for a solid, entropy as for a gas—is a good one.

It is more difficult to form an estimate of the agreement with experiment achieved in Eyring's work, since he does not make use of the interatomic forces determined in other ways, but, as we have seen, introduces an atomic radius, which in certain cases is allowed to vary with temperature; he is more concerned with the derivation of a correct partition function from experiment interpreted in the light of his model, than with the deduction of the properties of a liquid from those of the molecules of which it is built up. We wish, however, particularly to criticize the deduction of the entropy of fusion on p. 899 of reference (4). It will be seen that for the alkali metals ΔS is less than R, whereas according to Eyring's theory the liquid will always have an entropy of disorder R not possessed by the solid. Hirschfelder and Eyring seek to ascribe the discrepancy to an *increase* in the frequency of atomic vibrations on melting, which, by holding the atoms more rigidly in their mean positions, would decrease the entropy. Any increase in the frequency in a substance which expands on melting seems, however, to us to be extremely unlikely.*

§ 5. DEGREE OF DISORDER IN A LIQUID

We now turn to the theories which treat in greater detail the degree of order in a liquid; as we have seen, it is especially important to do this in the neighbourhood of the melting point, and at temperatures and pressures far from the critical point, where the similarity between solid and liquid is greatest.

A useful method of characterizing the internal configuration of liquids is by

* Hirschfelder and Eyring state that results on solid lithium show an increase of Θ with T. This is of course true, but is due to the replacement of the true frequency spectrum by the simplified Debye form. Actually the frequencies of all the normal modes of a solid decrease with increasing temperature.

means of the distribution function $g(r)$, which is a measure of the statistical density of molecules at a distance r from a given molecule; $4\pi r^2 g(r)\,dr$ is defined as the probability of finding a molecule between the distances r, $r+dr$ from a given molecule. For a simple liquid the x-ray scattering curve is related to the distribution function by Prins's equation

$$G(s)=\int_0^\infty 4\pi r^2 g(r)\frac{\sin(sr)}{sr}\,dr,$$

where $G(s)$ is the fraction of the incident x-rays scattered at an angle θ, and $s=\pi\sin(\frac{1}{2}\theta)/\lambda$.

A method of obtaining the distribution function for liquids *near the melting point* has recently been given by Prins[11]. Making use of the fact that the scattering curve of a liquid is, so to speak, a blurred copy of that of the corresponding crystalline solid, Prins has constructed distribution functions by allowing the integral numbers of molecules n_1, n_2, n_3 at fixed intermolecular distances r_1, r_2, r_3 characteristic of the crystalline solid to be displaced at random from their mean positions by distances Δr_1, Δr_2, ...; but in contradistinction to the state of affairs in a solid, where the thermal motion will of course give such a displacement, Δr_k is assumed to increase with r_k, so that the surrounding atoms get more and more out of phase with the semi-crystalline configuration in the centre as one goes away from it. Prins takes

$$\Delta r_k=\sqrt{(2\chi r_k kT)}/a,$$

where χ is the compressibility, and thus gives a measure of the restoring force of the atoms, and a is a numerical constant characteristic of the physical nature of the liquid. The distribution function is thus

$$g(r)=\sum_k g_k(r),$$

where

$$g_k(r)=\frac{n_k}{4\pi r_k^2}\sqrt{\left(\frac{a^2}{2\pi\chi r_k kT}\right)}\exp\frac{-a^2(r-r_k)^2}{2\chi r_k kT}.$$

In this way Prins shows how, to every principal type of crystal structure, there corresponds a liquid configuration, and how the distribution functions obtained for these give x-ray scattering curves in satisfactory agreement with experiment.*

Bernal[9], in an illuminating discussion of order in liquids, has generalized Prins's ideas in the following way. He assumes that liquids near the melting point are molecularly homogeneous, i.e. that there are neither large fluctuations of density nor large regions of regular quasi-crystalline structure separated by surfaces of misfit. These are not, of course, precise assumptions; they merely serve to direct one's train of thought in a certain direction. Bernal then attempts to define the distribution function in a liquid in terms of the following independent variables; the number of first nearest neighbours n_1; their mean distance from one another and from the central atom r_1, and the mean fluctuation λ of the interatomic distance

* For most liquids, simple and otherwise, there is considerable similarity between the x-ray diffraction patterns shown by the solid and liquid. The hypothesis of a cybotactic structure in liquids advanced by Stewart[12] in a number of papers to explain this fact is discussed below.

of these n_1 atoms about r_1. The distribution function of the n_1 first neighbours is taken to be

$$g_1(r) = \frac{n_1}{4\pi r_1^2} \sqrt{\left(\frac{1}{\pi r_1^2 \lambda^2}\right)} \exp\left\{-\frac{1}{\lambda^2}\left(\frac{r-r_1}{r_1}\right)^2\right\}.$$

By a semi-empirical, semi-geometrical method of investigating the ways in which spheres can be packed together, Bernal has shown that n_2, n_3, the numbers of molecules in the second and third co-ordination spheres, and their distances r_2, r_3, are roughly determined in terms of n_1, r_1 and λ; thus these three independent variables define the configuration of the liquid, at any rate if Bernal's assumption of the liquid's homogeneity be accepted.

He then attempts to write down the free energy F of the liquid. He takes a law of force

$$V(r) = \frac{a_m}{r^m} + \frac{a_n}{r^n}$$

between the atoms; the potential energy of the system will then be

$$\sum_k \int_0^\infty V(r)\, g_k(r)\, 4\pi r^2\, dr. \qquad \text{.....(12)}$$

He then adds a term $3kT \log (h\nu/kT)$ to take account of the vibrations, in our opinion incorrectly, the potential energy of the vibrations being already included in (12); since however ν is taken to be independent of n_1 and λ, this term is without influence on his final results. Finally he adds a term $-\phi T$, where ϕ is the entropy due to the configuration alone; it is not at present possible to calculate this; he assumes that it is proportional to λ^3. This appears to replace the constant term $k \log (N^N/N!)$ in the theory of Eyring.

Bernal then writes down the conditions that the free energy shall be a minimum for changes of λ and n_1. From the resulting equations he can obtain an idea of the variation of λ and n_1 with temperature. He finds that, as the temperature is raised, λ increases, as one might expect, and also that n_1 tends to decrease; as the temperature is raised the liquid becomes less close packed. Thus even in the simplest liquids we expect that the specific heat at constant pressure will contain a term representing the heat required to effect this change of n_1 in addition to the normal term representing the energy of the vibrations. We are thus led to the idea of a configurational specific heat in liquids, which will account for observed greater specific heats of simple liquids near the melting point compared with those of the solid.* The theory as it stands does not however account for the fall in C_p from a value above $3R$ at the melting point to a value round $2R$ at the critical point; this is certainly due to a fall in the vibrational energy, as the potential hole in which each atom moves changes from the form shown in figure 1 (*a*) to that shown in 1 (*b*) or 1 (*d*). Calculations along these lines have been made by Bartholomé and Euken[14].

Finally it must be stated that Bernal's method of attack has not reached the stage when numerical results can be obtained and compared with experiment.

* The idea of a configurational specific heat in liquids was actually given much earlier, cf. Simon[13].

Its interest lies in the attempt to fit a configurational co-ordinate into a quantitative theory.

A further point has been discussed by Bernal in the paper referred to. He asks whether it is possible to construct a series of distribution functions forming a continuum between an ordered and disordered state. Again by constructing empirically a number of aggregates of spheres, he concludes that it is not; for purely geometrical reasons there is no *homogeneous* intermediate state between a crystalline solid and a disordered liquid. The validity or otherwise of this statement must however depend on the use of the word homogeneous. Obviously a polycrystalline material is intermediate between a perfect crystal and an amorphous or liquid structure. Stewart is of the opinion that most organic liquids are best pictured as small crystalline aggregates separated by surfaces of misfit; such a hypothesis (the hypothesis of a "cybotactic" structure) accounts well for the observed X-ray diffraction patterns. Whether such a hypothesis gives the best description of liquid structure, or whether it is possible to form a homogeneous amorphous structure retaining the general features of the solid in some such way as described in § 6, seems to us a problem in solid geometry that has not yet been solved. Therefore, Bernal's conclusion, even if it could be established rigorously, does not entirely explain why crystalline solids melt at a definite temperature, and do not gradually become more disordered as the temperature is raised. On this point we shall now give some speculations due to the present authors[17].

§ 6. A LIQUID AS THE LIMITING FORM OF A POLYCRYSTALLINE SOLID

The recent work of Wagner[15], Jost[16] and others has shown that in crystalline solids in thermal equilibrium a certain degree of disorder exists at high temperatures, in the sense that a certain proportion of the atoms or ions of which the crystal is built up leave their normal lattice positions and move into the so-called interstitial positions. Disorder of this type is quite different from liquid disorder, since it does not destroy the long distance order of the lattice, or, in other words, the power to give sharp X-ray reflections.

In order to conceive of a gradual transition from the solid to the liquid, we shall imagine a perfect crystal gradually broken up into a polycrystalline mass of gradually decreasing crystal size. We do not wish to suggest that in an actual liquid there exist more or less perfect crystals separated by surfaces of misfit, or that such a hypothesis is necessary in order to account for X-ray results. We are here interested in the entropy of a liquid, and suggest only that for the estimation of the entropy a useful approach may be to treat a liquid as the *limiting* state of a polycrystalline solid in which the distinction between the individual crystals has disappeared[17]. Consider then first a polycrystalline mass in which the number n of atoms in each perfect crystal is fairly large. For the degree of order we define a variable λ equal to $n^{-\frac{1}{3}}$. The energy U of the aggregate will be proportional to the total area of the surfaces of misfit, and thus to λ. We set U equal to $A\lambda$. We discuss below the

definition of disorder when n is no longer large in comparison with unity, as is the case in the actual liquid.

The main difficulty of the problem is to estimate the configurational entropy due to the substance being a polycrystalline mass, and not a single crystal. We have to estimate the number of ways, P, in which a polycrystalline mass can be put together with energy $A\lambda$. As a very rough estimate, we shall assume that, if any individual microcrystal be rotated about any axis, the number of positions which it can assume is of the same order as the number of atoms on any line drawn round its surface. We thus have

$$P=[\alpha n^{\frac{1}{3}}]^3=[\alpha/\lambda]^3, \qquad \text{.....(13)}$$

where α is of the order unity, and for the entropy per gramme-atom

$$S=\frac{R}{n}\log\left(\frac{\alpha}{\lambda}\right)^3$$

$$=2R\lambda^3\log\left(\frac{\alpha}{\lambda}\right). \qquad \text{.....(14)}$$

We shall assume the entropy due to the lattice vibrations to be independent of whether the substance is polycrystalline or not.

Formula (14) certainly represents only a rough estimate of the entropy, but, the logarithmic term varying only slowly with T and n, we may from it deduce that the entropy is roughly proportional to B/n or $B\lambda^3$, where B is a constant. Thus, terms independent of λ being neglected, the total free energy $U-TS$ will be given by

$$F=A\lambda-BT\lambda^3.$$

We now consider what happens to the free energy F when λ becomes large, i.e. when the number of atoms in each crystal is no longer large, and when no distinction can be made between each perfect crystal and the surfaces of misfit. We can no longer increase the entropy by subdividing each crystal and putting together the bits at random; we can only increase the entropy by pushing each molecule out of its mean position in the field of its neighbours, or by increasing the volume of the substance. We may therefore expect the energy U to increase more rapidly than the entropy. If we define λ in such a way that the entropy remains proportional to B/n, the energy will vary with λ as in figure 5(*a*). The free energy for a series of temperatures such that $T_3>T_2>T_1$ will appear as in figure 5(*b*). For sufficiently high temperatures the free energy curves have a minimum. T_2 represents the temperature of melting.

At temperatures near the melting point a polycrystalline mass has no rigidity, since the boundaries between individual crystals can move freely through the body. In our view the fluidity of a liquid is due only to the presence of a large number of such boundaries. Below the melting point the free energy has its lowest value if the number of such boundaries decreases; the substance will recrystallize. Above the melting point, on the other hand, the free energy has its lowest value when the number of boundaries has a certain very large value.

In a close packed liquid, such as one of the liquid metals, it is a not unreasonable assumption that the individual crystal which can have any orientation relative to its neighbours consists of one atom and its twelve neighbours.

It may be that in the liquid phase there is some rotation of these groups. This idea could perhaps be related to a theory of Brillouin[18], according to which the degrees of freedom in a solid which are represented by *transverse* elastic waves are replaced in the liquid by whirl systems, which have no potential energy but only kinetic energy. According to Brillouin this leads to a thermal energy $\frac{1}{2}kT$ instead of kT per degree of freedom (except for the compressional waves). Brillouin accounts in this way for the drop in the specific heat as the temperature is raised above the melting point.

A physical way of estimating the disorder of a liquid metal is from the increase in its electrical resistance. It does not seem possible to relate this in any way to the concept of the sharing of the free volume. Before the publication of the papers

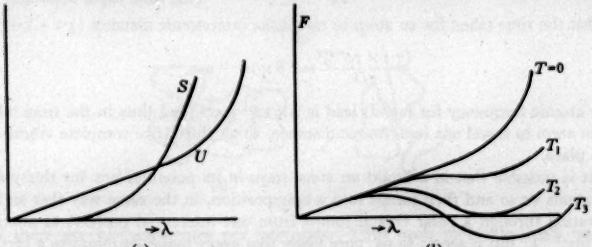


Figure 5. Entropy S, energy U and free energy F as functions of the disorder λ.

reviewed here, Mott[19] assumed that the change in the entropy of disorder on melting was negligible, and that the observed change L_i/T was due to the increase in "free volume" resultant on the breakdown of the crystalline form; actually the increase in free volume was expressed in terms of a decrease in the characteristic frequency ν. At a given temperature, therefore, the atoms would be further displaced from their mean positions in the liquid than in the solid. Since the electrical resistance in metals is proportional to the square of this displacement, the increase in resistance could be calculated in this way, the results being in good agreement with experiment. Although the present authors now consider that to treat the whole increase of entropy in this way is an over-simplification, the model does allow two quantities, the latent heat L and the increase in resistance ΔR, both of which depend on the disorder in the liquid, to be related.

Finally, a few words may be said about Lindemann's melting point formula, according to which, for metals, the atomic frequency ν and the melting point T_M are connected by the formula

$$T_M = \text{const. } m\nu^2\sigma^2,$$

where σ is the interatomic distance.

$2\pi^2 m\nu^2 x^2$ is a measure of the work done to displace an atom a distance x from its mean position. Thus we expect the work necessary to bring a substance into a given geometrical degree of disorder to be proportional to $m\nu^2\sigma^2$. Thus any theory of melting such as the one proposed above suggests that L and hence T_M should be roughly proportional to this factor.

§ 7. VISCOSITY AND DIFFUSION

Information about the actual paths followed by the molecules in a liquid can be obtained from data on diffusion.

The coefficient of self-diffusion of liquid lead near the melting point has been measured by Groh and von Hevesy[20], who observed the diffusion of radioactive lead (ThB) in the non-radioactive liquid metal. The coefficient D is given by

$$D = 2{\cdot}55 \times 10^{-5} \text{ cm.}^2/\text{sec.},$$

so that the time taken for an atom to travel the interatomic distance ($3{\cdot}1$ A.) is

$$\frac{(3{\cdot}1 \times 10^{-8})^2}{2D} = 1{\cdot}8 \times 10^{-11} \text{ sec.}$$

The atomic frequency for (solid) lead is 2×10^{12} sec.$^{-1}$, and thus in the time taken by an atom to travel one interatomic distance, about thirty-five complete vibrations take place.

It is possible that in a liquid an atom stays in its potential box for thirty-five vibrations or so and then jumps into a new position, in the same way that an ion migrating through a polar crystal jumps from one interstitial position to another (cf. Jost[16]). But it seems to us more likely that every molecule migrates a certain distance during each vibration; this distance may be estimated by making use of a theorem of Lord Rayleigh, according to which if n consecutive paths of length ξ succeed one another in random directions, the probable total displacement is

$$\sqrt{\tfrac{3}{8}}\,\xi n^{\frac{1}{2}}.$$

Equating this to the interatomic distance σ, and n to the number 35 found above, we obtain

$$\xi = 0{\cdot}2\sigma.$$

Now in a solid, if an atom has the mean thermal energy $3kT$, then at any moment when it is at rest its displacement x from its mean position is given by

$$3kT = 2\pi^2 m\nu^2 x^2. \qquad \text{.....(15)}$$

According to the empirical formula of Lindemann, x calculated at the melting point from this formula is about $0{\cdot}09\sigma$, where σ is the interatomic distance. Thus in a solid near the melting point each molecule is performing vibrations as shown in figure 6 (*a*) covering a length of the order $0{\cdot}18\sigma$. In the liquid then, since this is equal to ξ, in order to obtain the required diffusion coefficient, we must imagine that *each* vibration takes place in a direction unrelated to the previous one, figure 6 (*b*).

We now review certain recent theoretical work on viscosity. Andrade[21] has found that the viscosity of a number of metals at the melting point is given approximately by the formula

$$\eta = \frac{4}{3}\frac{m\nu}{\sigma},$$

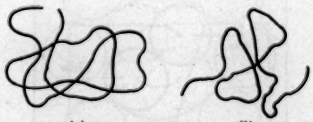
where σ is the average distance between the centres of molecules, m the molecular mass, and ν the frequency of vibration for the corresponding solid.

Andrade interprets his formula in the following way: he considers two successive layers of atoms in the liquid sliding over one another with relative velocity $\sigma\partial v/\partial z$, where v is the velocity perpendicular to the z direction. Each molecule then has a mean momentum

$$\sigma m\, \partial v/\partial z$$

relative to the layer below it. Now the momentum transferred across the layer per unit time is per unit area

$$\eta \frac{\partial v}{\partial z},$$



(*a*) (*b*)

Figure 6.

and there are $1/\sigma^2$ molecules per unit area. Thus each molecule gives up its excess momentum to the layer below it n times per second, where

$$n = \sigma^2\eta \frac{\partial v}{\partial z} \Big/ \sigma m \frac{\partial v}{\partial z} = \frac{\sigma\eta}{m}.$$

Now we have seen that in a liquid near the melting point a molecule changes its direction of motion once in a time of the order ν; thus it seems reasonable to assume that n and ν are of the same order, so that

$$\eta = \text{const.}\, \frac{m\nu}{\sigma},$$

where the constant is of the order unity.

Eyring[5] considers that the motion of one layer of liquid over another can only take place at points where, owing to the thermal fluctuations in the density of the liquid, the free volume available is particularly high. The type of motion envisaged is shown in figure 7. The essential difference between Eyring's point of view and Andrade's, then, is that while according to Andrade all the atoms of a simple liquid under flow are acquiring momentum and passing it on to the other layers *all the time*, according to Eyring at any moment only some of the atoms can move. Eyring thus obtains for the viscosity

$$\eta = Ae^{-W/kT}, \qquad \text{.....(16)}$$

where W is the activation energy for the formation of a "hole", or region of low density, and A is of the same order of magnitude as Andrade's factor $\sigma\eta/m$. Eyring points out that if a hole of volume v/n is required, where v is the atomic volume, then W will be equal to W_0/n, where W_0 is the heat of vaporization per molecule. He arrives at this result in the following way: to take a molecule from the middle of a liquid or solid to a point outside it requires energy $2W_0$; to add it to the outside requires energy $-W_0$. He then extrapolates this result to holes of less than one atomic volume.

For argon, N_2 and a number of organic non-associated liquids in which the molecules are held together by van der Waals's forces, Eyring points out that the observed viscosity is given by a formula of the type (16), with $W/W_0 \sim \frac{1}{3}$.

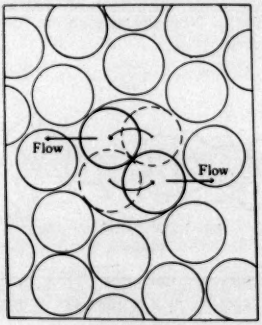


Figure 7.

This gives an indication of the size of the "hole" necessary for flow to take place. For liquid metals on the other hand, W/W_0 ranges from $\frac{1}{8}$ to $\frac{1}{30}$. Eyring considers that this shows that it is only the ions which flow in metals, and since the ion is much smaller than the atom a much smaller hole is necessary than one would expect from the atomic diameter. We do not like this explanation. It would suggest that, in comparison with the lattice energies, the forces in metals keeping the ions in their mean positions are abnormally small. This would mean that the atomic frequencies ν and hence the characteristic temperatures Θ in metals were abnormally small compared with the lattice energies. If a comparison between $mo^2\nu^2$ and W_0 is made for a number of substances, there is no very great difference between metals and non-metals[17].

We think it much more likely that, if Eyring's theory of holes has any significance, the holes are of the same size as in non-metals, but that the energy necessary to form them is much less. It is known from the work of Jost[16] that the energy necessary to form a hole in an ionic crystal is much less than W_0, because when

an *ion* is removed from a crystal the medium becomes polarized round the hole, and the energy of polarization compensates to some extent for the work necessary to remove an ion. Similar calculations have not been carried out for metals, but an effect of the same kind may very well occur. If, as Wagner[27] has suggested, diffusion in solid metals is due to the presence of holes, the activation energy for their formation must be less than W_0.

§ 8. LIQUID HELIUM

The properties of helium will be discussed in this article only in so far as they bear on the general problem of liquid structure. As shown in figure 8, liquid helium exists in two forms, known as liquid He I and liquid He II. At pressures below 24 atmospheres helium remains a liquid down to 1° K. and presumably to the absolute zero of temperature. Any substance in thermodynamical equilibrium at the absolute zero must have zero entropy. There are thus two possibilities:

(i) That what is called liquid helium is a crystalline solid which for some reason has no rigidity.

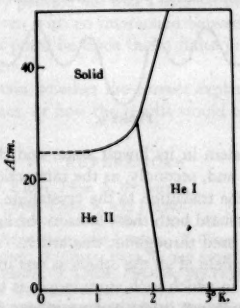


Figure 8.

(ii) That the atoms in liquid helium are disordered as in an ordinary liquid, but that disorder and zero entropy are not necessarily incompatible.

F. London[22] has expressed the first point of view. He has given evidence to show that, owing to the very high zero point energy of liquid helium, the diamond structure would be the most stable, since it gives to each atom the greatest free volume. He suggests that such a structure might have zero rigidity.

Since the publication of London's paper, however, Keesom[23] and his collaborators have shown that liquid helium at 1·6° K. gives broad X-ray diffraction rings like an ordinary liquid or compressed gas; it now appears more likely, therefore, that even at the absolute zero of temperature liquid helium has no crystalline structure.

If there were no interaction between the atoms, an assembly of N atoms at the absolute zero of temperature in an enclosure of volume V would form a fully degenerate Einstein-Bose gas. Each atom would be represented by a wave function $\psi(r)$ and the wave function of the whole system be

$$\psi(r_1)\,\psi(r_2)\ldots,$$

ψ being the wave function for a particle in its lowest state in the enclosure; if we take the enclosure to be a rectangular box of sides a, b, c,

$$\psi(r) = \sin\frac{\pi x}{a}\sin\frac{\pi y}{b}\sin\frac{\pi z}{c}.$$

In such a gas there is, of course, no crystalline structure, and in the lowest state the entropy is zero. The question now arises whether, if we introduce an interaction between the atoms, still keeping the system in its lowest quantized state, i.e. at zero temperature, we *necessarily* go over to a crystalline structure, or whether, if the interaction is not too large, the system will remain amorphous, like an imperfect gas. It has not yet proved possible to deduce from quantum mechanics an answer to this question; but the properties of liquid helium seem to provide an experimental

Figure 9.

answer, namely, that a system in its lowest state, and thus with zero entropy, is not necessarily crystalline, and, secondly, as the interaction is increased (i.e. as the substance is compressed) the transition to the crystalline state is sharp.

We may perhaps understand both these facts in the light of the Einstein model of a solid which has been used throughout this article. In this model, each atom is supposed to move in the field of all the others at rest in their positions of equilibrium. The potential field in which each atom moves is thus as shown in figure 9. The subsidiary hollows represent interstitial positions; for instance, in the face-centred cubic structure these would be the centres of cubes. As the temperature is raised, a few atoms leave their normal positions and move into interstitial positions (as happens for ions in Wagner's[15] description of ionic conduction in solid polar salts). In the potential hole which represents an atom's normal position there are normally a whole series of stationary states; but if the mass of the atom or the depth of the hole are small enough, there will not be a single stationary state in the hole;* the wave function of an atom in its lowest energy state spreads through the crystal. Under these conditions it is clear that no solid is possible; and the criterion, that the potential hole must be deep enough to contain at least one stationary state if a solid is to be possible, gives a sharp limit beyond which the solid cannot exist.

* In the one-dimensional problem a potential hole always contains at least one stationary state, but this is not the case in the three-dimensional problem.

The hypothesis that liquid He II is to be regarded as a (very imperfect) degenerate Einstein-Bose gas seems to us tenable. We shall now see to what extent it is capable of accounting for the thermal properties of the substance.

§ 9. THE λ POINT

Near 2° K., at temperatures shown by the line in figure 8, the specific heat of liquid helium shows a kink. At temperatures and pressures above this line the liquid (liquid He I) has normal properties; below it, on the other hand, the liquid (He II) shows abnormal properties (low viscosity, high thermal conductivity). The critical temperature is known as the λ point.

Assuming liquid He II at the absolute zero to be a crystalline solid with the diamond structure, Fröhlich[24] has supposed that as the temperature is raised atoms move into interstitial positions (as in Wagner's theory of ionic conductivity). The work necessary to move an atom into an interstitial position is shown to decrease with the number already moved. The process is thus similar to the transition from the ordered to the disordered state in β-brass, recently treated by Bragg and Williams, and a sharp kink of the observed kind is to be expected.

London[25] has recently pointed out that a degenerate Einstein-Bose gas should show a transition point, even with no interaction between the atoms. At temperatures below the transition point he finds that a finite proportion of the atoms are in the lowest quantum state.

It is not yet quite certain whether the correct explanation of the phenomenon is to be found on these lines, or how the results would be altered if the interaction were taken into account.

REFERENCES

(1) Mayer, J. E. *J. Chem. Phys.* **5**, 67 (1937).
(2) Eyring, H. *J. Chem. Phys.* **4**, 283 (1936).
(3) Eyring and Hirschfelder. *J. Phys. Chem.* **41**, 249 (1937).
(4) Hirschfelder, Stevenson and Eyring. *J. Chem. Phys.* **5**, 896 (1937).
(5) Ewell and Eyring. *J. Chem. Phys.* **5**, 726 (1937) (Viscosity).
(6) Kinkaid and Eyring. *J. Chem. Phys.* **5**, 587 (1937) (Mercury).
(7) Lennard-Jones. *Physica*, **4**, 941 (1937).
(8) Lennard-Jones and Devonshire. *Proc. Roy. Soc.* A, **163**, 53 (1937).
(9) Bernal. *Trans. Faraday Soc.* **33**, 27 (1937).
(10) Prins. *Z. Phys.* **56**, 617 (1929).
(11) Prins. *Physica*, **3**, 147 (1936).
(12) Stewart, G. W. *Phys. Rev.* **46**, 706 (1934).
(13) Simon. *Ergebn. exakt. Naturw.* **9**, 248 (1930).
(14) Bartholomé and Euken. *Trans. Faraday Soc.* **33**, 45 (1937).
(15) Wagner and Beyer. *Z. phys. Chem.* B, **32**, 113 (1936).
(16) Jost. *Diffusion und Chemische Reaktionen in festen Stoffen* (Leipzig, 1937).
(17) Mott and Gurney. *Trans. Faraday Soc.* (in the press).
(18) Brillouin. *Trans. Faraday Soc.* **33**, 54 (1937).
(19) Mott. *Proc. Roy. Soc.* A, **146**, 465 (1934).
(20) Groh and v. Hevesy. *Ann. der Physik*, **63**, 85 (1920).
(21) Andrade. *Phil. Mag.* **17**, 497, 698 (1934).
(22) London. *Proc. Roy. Soc.* A, **153**, 576 (1936).
(23) Keesom and Taconis. *Physica*, **4**, 256 (1937).
(24) Fröhlich. *Physica*, **4**, 639 (1937).
(25) London. *Nature, Lond.*, **141**, 643 (1938).

PLASTICS IN INDUSTRIAL PHYSICS

By H. W. MELVILLE

Colloid Science Laboratory, Cambridge

THE term "plastics" covers such a large variety of materials that it is first necessary to define more precisely what is meant by the word in order to see at what points contact is made with industrial physics. The word rather denotes a state of matter than a group of chemically well-defined substances. Accordingly it will be pertinent to review briefly what physical conceptions are now extant concerning their structure and how far such theories account for the peculiar, and as yet imperfectly understood, bulk phenomena exhibited by plastics.

Plastics owe their characteristics mainly to the fact that the constituent molecules are of high molecular weight. The shape of the molecules also plays a very significant part in modifying mechanical and electrical properties. Before any progress can be made in accounting for and utilizing their behaviour to the best advantage, the size and structure of the molecules must be known with certainty. What is of not less importance is that both size and structure must be capable of alteration in a predetermined manner. For the most part we shall therefore exclude natural plastics and concentrate on the synthetic product, since it is here that the significant advances have been or are likely to be made in systematizing and clarifying the complex phenomena involved.

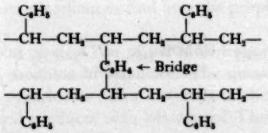
The efforts which are being made in this latter direction are therefore bound up intimately with the modes of synthesis. Attention must first be directed to the various types of synthetic process used for the manufacture of these resins. Essentially such methods may be divided in two groups (*a*) true polymerization, (*b*) condensation. In the first method the monomeric molecule which forms the polymer interacts with another of its kind with the formation of a so-called dimer. Another molecule of monomer adds on and so on, a linear polymer being produced thus:

$$CH_2{=}CH_2 + CH_2{=}CH_2 \rightarrow CH_3{-}CH_2{-}CH{=}CH_2$$
$$+ n\,(CH_2{=}CH_2) \rightarrow CH_3{-}CH_2{-}(CH_2{-}CH_2)_n{-}CH{=}CH_2.$$

Alternatively instead of the polymer having a double bond at the end, it may possess a structure of the following kind: $X{-}(CH_2{-}CH_2)_n{-}Y$, where X and Y are groups which originally functioned as catalyst and inhibitor of chain growth respectively. This simple type of reaction illustrates well the point made in the introductory remarks about the size of the molecule very largely determining its behaviour. If ethylene is induced to polymerize by compressing to pressures of about 1,000 atmospheres, the value of the n is so large that the product, instead

of having the familiar waxy properties of short hydrocarbon chains ($n = 15$ to 20), has elastic properties resembling rubber.

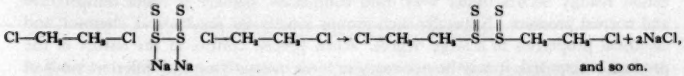
Ethylene is, however, rather difficult to polymerize to molecules of high molecular weight, but this may be easily overcome by substituting for one or both hydrogen atoms on the carbon atom of the double bond various groups such as Cl, CN, $COOCH_3$, $—O.CO.CH_3$, (the phenyl radical), when polymerization readily occurs under very mild conditions, namely at room temperature and normal pressure. Naturally such groups modify the mechanical, chemical and electrical properties to a large degree. When precise control of the nature of the product is required, it may be necessary to form interpolymers of different kinds of monomer. This type of reaction, however, only gives rise to long chain polymers. For some purposes the chains may have to be linked together by bridges. This may also be accomplished by using a compound containing two double bonds separated by a number of carbon atoms, for example, the compound (*p*-divinyl benzene) may cross-link chains formed by the polymerization of similar molecules, namely (styrene)[1]. In this way a three-dimensional network is built up thus:



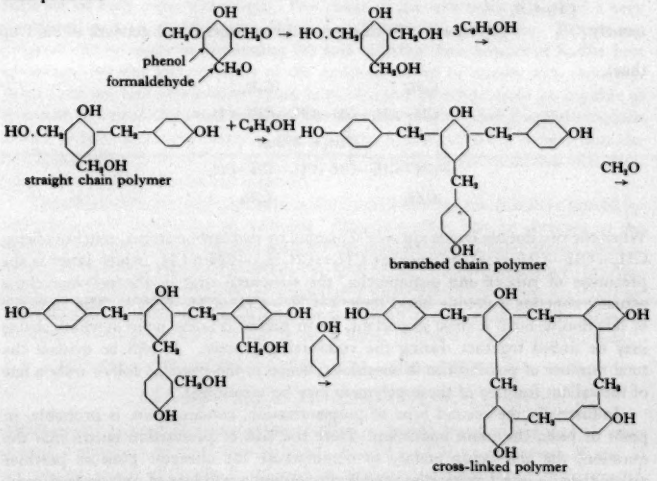
When the two double bonds are only separated by two carbon atoms, as in butadiene, $CH_2=CH—CH=CH_2$ or isoprene $CH_2=C(CH_3)—CH=CH_2$, which latter is the precursor of rubber and guttapercha, the structural unit in the polymer chain actually contains a double bond thus: $—CH_2—CH=CH—CH_2—$. The presence of this double bond is most important, for in rubber it is the point at which chains may be linked together during the vulcanizing process. As will be evident the total number of possibilities is enormous, hence in the pages to follow only a few of the salient features of these polymers may be mentioned.

Industrially the second type of polymerization, condensation, is probably, in point of bulk, the more important. Here the cost of production enters into the question, the aim being mainly to manufacture the cheapest possible polymer rather than to select materials capable of yielding a polymer of any desired composition. In the somewhat empirical state of the subject, this imposed choice has not militated against obtaining a suitable product. In polycondensations the essential feature is that a simple molecule separates each time a monomer reacts with the growing polymer. Unlike most syntheses in organic chemistry all the

steps leading to the formation of the large molecule take place in one reaction system. In fact it is this fortunate circumstance which makes the production of plastics a practicable proposition. A simple type of condensation is that between sodium tetrasulphide (Na_2S_4) and ethylene dichloride to form thiokol rubbers which possess the superiority over natural rubber in that they do not swell nor dissolve in hydrocarbon solvents. The reaction is as follows:

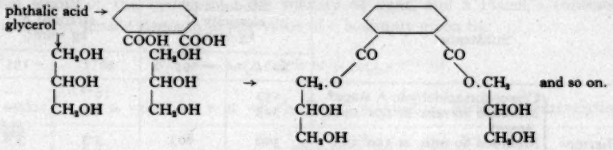


There are three further types of condensation reactions which deserve mention, because the products have found extensive application in industrial physics. The most important are those involving the interaction of phenols and aldehydes in presence of a condensing catalyst, of which bakelite is the best known example. The exact manner in which the units are linked together is not yet known with certainty, but the following scheme shows the nature of the reactions leading to the final product.



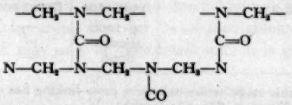
The first stage is the formation of a polyhydric alcohol, next the formation of a straight chain polymer, followed by a branched polymer and finally by a cross-linked polymer. All three processes probably occur more or less simultaneously during the condensation.

Another important group of resins is formed by esterifying polyhydric alcohols such as glycerol with polybasic acids such as phthalic acid:



Here it is evident that straight and cross-linked polymers will be formed. The name glyptal is attached to this series of products.

Finally mention may be made of the urea $CO(NH_2)_2$-formaldehyde resins. The chemistry is too complicated to be discussed here, but the final structure is probably extensively cross-linked thus:



These few examples will serve to show how a great variety of molecular types may be built up. It is now expedient to find how the properties of plastics depend on their constitution.

From the present point of view two of the most important physical properties are tensile strength and modulus of elasticity. The question immediately arises, Can the resins be made superior to other materials in these respects so that they might replace such natural products with advantage? The answer is for the most part in the negative, although, as will be shown later, there are specific requirements which can only be met by employing synthetic resins. A glance at Table 1 [a] shows immediately that both modulus of elasticity and tensile strength do not exceed the values for other substances potentially replaceable by resins.

As will be evident from the figures for the thermosetting resins a very marked increase in both quantities occurs when further polymerization, including chain lengthening and cross-linking, takes place. A maximum is soon reached which appears to be an upper limit for this type of plastic.

Similarly thermoplastic polymers possess values of E and of F of the same order of magnitude while spring steel heads the list, though many cellulose derivatives, having regard to their density, are not far behind. In order to see whether it is worth while to attempt to improve on these figures, calculations have been made of the maximum of the possible values of E and F. This can be done from two extreme points of view. It may be assumed that the chains are so long that when any stress is applied to the solid, the resulting strain is due to the displacement of atoms joined together by primary valencies either by deformation or by stretching. Actually chains are very much shorter than the dimensions of the specimen under

Table 1

Substance		Modulus of elasticity E kg./mm.^{2}		Tensile strength F kg./mm.^{2}	
		20° C.	−195° C.	20° C.	−195° C.
Thermo-setting resins	Cresol-formaldehyde A stage*	152	—	10^{-2}	—
	Moulded 10 min. at 150° C. (C stage)	513	—	2·5	—
	Moulded 60 min. at 150° C.	500	803	3·3	3·8
	Phenol-formaldehyde A stage*	290	—	10^{-2}	—
	Moulded 10 min. at 150° C. (C stage)	591	—	6	—
	Moulded 60 min. at 150° C.	595	1050	6·7	7·8
	Urea-formaldehyde (C stage)	350	—	3	—
Thermo-plastic resins	Aniline-formaldehyde	300	—	7	—
	Polystyrene, mol. wt. 48,000, straight chain	383	452	4·1	4·4
	Cellulose products	350–2200	650–4200	20	10–12
	Glass†	4–1000	—	3·5–8·5	—
	Spring steel	20000	—	195	—
	Duralumin	8000	—	36	—

* That is, fusible or soluble stage before extensive cross-linking has occurred.
† A. A. Griffith, *Philos. Trans.* A, **223**, 163 (1921).

A thermosetting resin is one which may be moulded to the required shape in the early stages of polymerization, but when polymerization is carried to completion, the product cannot then be induced to flow at any temperature. On the other hand moderate heat treatment of thermoplastic substance causes it to soften and if necessary to flow into any desired shape, which is retained on cooling.

test. For example, a molecule composed of 10^5 units (e.g. a molecular weight of 10^7, assuming a molecular weight of the monomer to be 10^2) would be considered very large, yet the chains when fully stretched would only be about 10^{-2} cm. long. Consequently there are experimental difficulties in carrying out the test under conditions imposed by the assumptions made in working out a theoretical expression for both modulus and tensile strength. Alternatively the assumption may be made that the chains are short and are orientated parallel to each other and to the direction in which the stress is applied. When the stress is applied the chains slip past each other, the force opposing the motion being of the van der Waals type due to the interaction between the chains themselves. J. H. de Boer[3] has carried out such calculations. The first kind of calculation is the easier to carry through. As a first approximation, the problem is to calculate the tensile strength and Young's modulus of a single C—C bond. The potential energy (Q) of such a bond as a function of the internuclear separation r (r_0 is the equilibrium separation) is given by the empirical Morse equation

$$Q = Q_0 e^{-2a(r-r_0)} - 2Q_0 e^{-a(r-r_0)},$$

where Q_0 is the energy of dissociation and a is given by the equation

$$a = \sqrt{(8\pi^2 c \mu \omega_0^2 / h)},$$

where μ is the reduced mass, ω_0 is the frequency of the interatomic vibration, x the anharmonic factor for the vibration and calculable from the spacing of the vibrational levels of the molecule, c the velocity of light, and h Planck's constant. Hence F the tensile strength at any value of r is simply given by

$$F = \frac{\partial Q}{\partial r} = 2aQ_0 e^{-2a(r-r_0)} + 2aQ_0 e^{-a(r-r_0)}.$$

In order to get a value for r at which F is a maximum, further differentiation yields

$$r_{\max.} = \frac{ar_0 + \ln 2}{a}.$$

Taking $a = 1{\cdot}0 \times 10^8$ cm.$^{-1}$ and $r_0 = 1{\cdot}54 \times 10^{-8}$ cm., $r_{\max.} = 2{\cdot}23$ A. Hence

$$F\,(\max.) = 2aQ_0\{e^{-\ln 2} - e^{-2\ln 2}\} = \frac{aQ_0}{2} = 5{\cdot}64 \times 10^{-4} \text{ dyne.}$$

The estimated tensile strength for phenol-formaldehyde resin, assuming 8×10^{14} bonds per cm.^{2}, is 4300 kg. mm. That for cresol-formaldehyde is 3800 kg. mm.$^{-2}$

Young's modulus E may now be calculated since it is given by

$$E = \left(\frac{\partial F}{\partial r}\right)_{r=r_0} \times r_0 \times (1 - 2m),$$

where m is Poisson's ratio, which is probably about 0·3[4]. Hence $E = 35{\cdot}8 \times 10^{-4}$ dyne per bond. With $7{\cdot}6 \times 10^{14}$ bonds per cm.^{2}, $E = 11{,}000$ kg. mm.$^{-2}$ The above calculations are subject to some uncertainty owing to the fact that the value of a is not known exactly for the system under consideration, but the result is certainly correct to within less than an order of magnitude.

The calculation of the van der Waals interaction is a very much more difficult matter and the argument of de Boer[5] too long to be discussed profitably in this article. It will suffice merely to quote the results of the calculations (table 2).

Table 2

Substance	Tensile strength (kg./mm.^{2})		
	Primary valency	van der Waals	Observed
Phenol-formaldehyde	4300	39	7·8
Cresol-formaldehyde	3800	32	3·8
Sodium chloride	400	20	0·6
	Young's modulus (kg./mm.^{2})		
Phenol-formaldehyde	11000	45	595

The fact that the observed tensile strength is less even than that due to van der Waals forces would imply that only such forces hold the solid together, but since the observed value of Young's modulus is between the two calculated values, the conclusion to be drawn is that primary valencies do play an important part in determining these mechanical properties. The discrepancy seems to be due to an overestimate of the number of effective bonds per mm.^{2}, for if the resin under test

is regarded as single molecule then it is not improbable that in a number of places there are far fewer primary valencies perpendicular to any selected plane than that computed from the number of atoms in the plane. Such an explanation is in a way analogous to that put forward to account for the very low tensile strength of sodium chloride as compared with the calculated value. With the thermosetting resins it is therefore unlikely that the above values will be bettered, but on the other hand it may be possible to grow long chain polymers with a controlled amount of cross-linking so that improved characteristics will be obtained. The next problem will be that of attaching such materials to a suitable anchorage by primary valencies in order that full use may be made of the exceptional properties. Some progress has already been made in this direction in that it has been found possible to bond rubber to metal so effectively—by what type of bond is as yet uncertain—and in such a manner that when the system is subjected to stress the break occurs within the rubber and not at the metal-rubber interface[6].

One of the most useful properties of certain plastics is the enormous extension which they may undergo without permanent deformation, the stress-strain curve exhibiting marked deviation from linearity. Sufficiently high stress, of course, leads to plastic deformation when the system behaves very much like a fluid. Besides being exhibited by natural rubber, such characteristics are possessed by many plastics, for example, those synthesized from units similar in structure to that found in natural rubber, namely butadiene, $CH_2{=}CH{-}CH{=}CH_2$, chloroprene $CH_2{=}CCl{-}CH{=}CH_2$, thiokols, polyethylene and even thermosetting resins under suitable conditions (though in the latter case the effect is rather transient)[7]. Temperature, however, has a most important effect in that highly elastic properties disappear at low temperatures (for example rubber may be powdered at liquid air temperatures). Similarly at high enough temperatures some polymers—polystyrene and polyvinyl acetate—become highly elastic[8]. This phenomena may be put to practical use for it may be desirable to ensure that high elasticity should be preserved at low working temperatures. For such a purpose a low "high elasticity temperature" is necessary. In this respect thiokol rubber is superior to natural rubber since high elasticity is maintained down to $-40°$ C.[9]

We need hardly enter into a discussion of the relative merits of the various theories for the explanation of high elasticity for these have already been discussed at length in various publications[10]. The attractive kinetic theory in which unstretched rubber is supposed to consist of long molecules coiled up into a ball-like structure and which on stretching assumes a fibrous structure is discussed by Meyer, Mark and others[11].

Having now given a brief survey of some of the general points concerning plastics it is now convenient to mention a few of the specific advances made within the past two or three years.

The properties of interpolymers. Much research is directed to the question of the production of interpolymers in order to obtain products having characteristics intermediate between those of simple polymers. For example vinyl chloride ($CH_2{=}CHCl$) and vinyl acetate ($CH_2{=}CH.O.CO.CH_3$) readily polymerize in

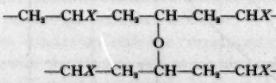
presence of a wide variety of catalysts. The former polymer is not thermoplastic to a high degree; it is also insoluble in most solvents and is not very stable to heat and light. On the other hand polyvinyl acetate softens at 30° to 40° C. but is very stable to heat and light and soluble in most solvents. An intimate mechanical mixture of the two polymers is not sufficient to produce a resin of intermediate properties. This can only be achieved by incorporating the monomer units in one and the same chain[12] when the desired properties are immediately obtained. For most purposes about 85 per cent of vinyl chloride is used to obtain an interpolymer which is thermoplastic, heat and light resisting and also resistant to chemical attack. If, however, a more plastic material is necessary for compatibility with say nitrocellulose so that the mixture may flow readily when employed for spreading in thin films, as a protective coating, then the vinyl chloride content may be reduced to 65 per cent. The properties of the interpolymer are also controlled by the molecular weight. Generally the latter is measured by a method originally due to Staudinger[13], who deduced, assuming that molecules are long single chains of atoms, that

$$\frac{\eta_s - \eta_0}{\eta_0} = KcM,$$

where η_s and η_0 are the viscosities of a solution of the polymer of molecular weight M and concentration c (g./litre) and of the solvent respectively. K is a constant. Its value is determined by choosing polymers of low molecular weight, the latter being measured directly by a cryoscopic method. The accuracy of the method for high polymers thus depends on rather a long linear extrapolation, but has been checked by Signer and Gross[14] for polystyrene by the ultracentrifuge with fairly good agreement[15].

Some results for the vinyl chloride-vinyl acetate polymer are shown in figure 1. It will be seen at once that there is no advantage to be gained in mechanical properties at least by increasing the molecular weight above 10,000, for at higher molecular weights, the polymer becomes unworkable on account of insolubility and infusibility.

Vinyl acetate resins may also be improved by cross-linking the chains by means of divinyl ether[16] ($CH_2{=}CH.O.CH{=}CH_2$) thus:



Both fusibility and solubility decrease as the proportion of divinyl ether is increased.

Hysteresis in plastics. When a plastic material such as rubber is subjected to stress a certain quantity of heat is generated with the result that when the stress is removed the stress-strain curve instead of retracing the path *OAB*, figure 2, follows *BCO*. The area enclosed by the two curves represents the heat dissipated. The question then arises as to how this area varies with the frequency of the application of the stress. This question is of practical importance in determining

the rate of tyre wear, the rubber being subject to such compressions some 10 times per second if a car is travelling at say 40 m.p.h.

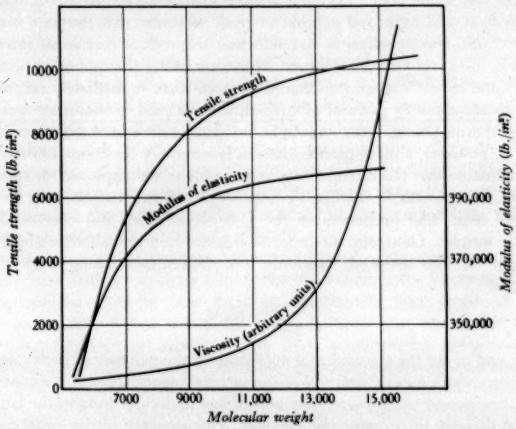


Figure 1. Variation of properties of vinyl chloride-acetate interpolymer with molecular weight.

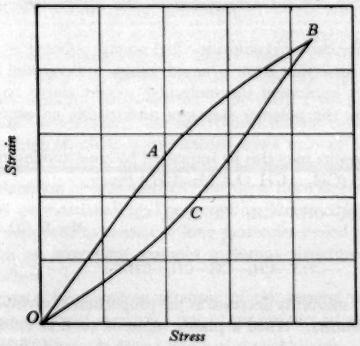


Figure 2. Hysteresis curve of a highly elastic substance.

Again with the wide-prevailing tendency to mount internal combustion engines on rubber, it becomes necessary to find when this mounting shows signs of fatigue, the period between successive applications of the stress being as small as 10^{-2} sec.

Machines for low frequency testing of rubber specimens have been employed for some time, but it is only recently that Naunton and Waring[17] have designed an apparatus for high frequency (up to 700 c./sec.) testing after the rubber has seen service in the above-mentioned connexions. It consists of two cylinders of rubber attached rigidly to either side of the moving coil of a permanent magnet loud-speaker. The coil is driven by an audio-frequency oscillator with an output of 30 w. which is sufficient for the purpose. If, however, it is also required to produce fatigue in the specimen, at least 10 times this output would be required. Arrangements are made for compressing the rubber and for loading the coil with weights so that the resonant frequency of the system can be adjusted in the range 0 to 700 c./sec.

The method of evaluating the elastic constants is best seen from the theory of the instrument. If the rubber is regarded as an ideal spring of spring constant s dynes/cm. and b is the coefficient of viscous resistance which is proportional to the velocity of displacement, then

$$m\frac{d^2x}{dt^2}+b\frac{dx}{dt}+sx=P\cos pt,$$

where m is the mass of the system, x the displacement, P the stress and p the angular frequency. Under the conditions employed, the approximate solution of the above equation is

$$x_{\max.}=\frac{P}{bp}\sin pt;$$

also

$$s=mp^2.$$

Since P, p and m are known and $x_{\max.}$ is measured, b and s may be computed. If E is the modulus of elasticity then

$$E=\frac{sL}{2\pi R^2},$$

where L and R are the length and radius of the block. The resilience ϕ is simply defined as the ratio of two successive amplitudes when the oscillation is free. This is given by the factor $e^{-bp/2m}$; the percentage is 100 times this figure.

The modulus of elasticity linearly increases with increasing frequency, the rate of increase depending on the nature of the rubber, its fillers and mode of vulcanization; it may amount to as much as a twofold increase from zero to 700 c./sec. On the other hand percentage resilience decreases in a like ratio, though there does not appear to be any connexion with the corresponding coefficient for the modulus of elasticity. Likewise the viscous resistance also exhibits a decrease with increasing frequency. This machine has already been used for preliminary experiments on rubber fatigue, for example, samples suitable for tyre treads were subjected to torsional fatigue for a period corresponding to 500 miles of hard use. The effect was to decrease the modulus of elasticity, though the variation of the modulus with frequency remained practically unaltered.

Electrical properties of synthetic resins. In the application of synthetic resins in the electrical industry a large number of factors have to be taken into consideration[18]. Besides the electrical characteristics, such as dielectric loss, dielectric constant,

breakdown voltage and insulation resistance, account must be taken of the mechanical properties of the resins and their stability to heat, light and the action of moisture and of oil. Thermosetting resins have been used for many years either mixed with a suitable filler and moulded under pressure or as an impregnant for fibrous materials such as cardboard. For some purposes, however, as for example cable sheathing, thermoplastic resins are much more suitable and promise to come into extensive use. The possible modes of application divide these resins into three groups (1) pure hydrocarbons, (2) oxygen containing polymers and (3) chlorine containing polymers. Typical examples are shown in the following table:

(1)	(2)	(3)
Polystyrene	Polyacrylate	Polyvinyl chloride
Buna	Polymethacrylate	Polychloroprene
Polyethylene	Polyvinyl acetate	

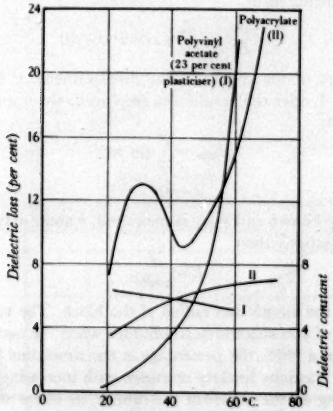


Figure 3. Dielectric loss at 50 c./sec.

The pure hydrocarbons are by far the best polymers if dielectric loss must be kept at a minimum. At room temperature and at a frequency of 800 c./sec., the loss is only 2 to 3×10^{-4} and is not much dependent on temperature, whereas as figure 3 shows the polyacrylates and polyvinyl chloride (with plasticizer) exhibit very high dielectric loss at the comparatively low frequency of 50 c./sec. Similarly cellulose triacetate has a high loss which however does not vary much with temperature in the range 20° to 80° C. The disadvantage of polystyrene is its solubility in mineral oil. In order to eliminate this defect it has been suggested that the polystyrene might be enclosed in an oil resistant resin, such as polyacrylate or polychloroprene. On the other hand if a water resistant insulation is required, the

acrylate resins cannot be employed, a plasticized polyvinyl chloride showing marked superiority in this respect. Some indication of the uptake of water and oil by such resins is given in figures 4 and 5. Whether it will be possible in future to

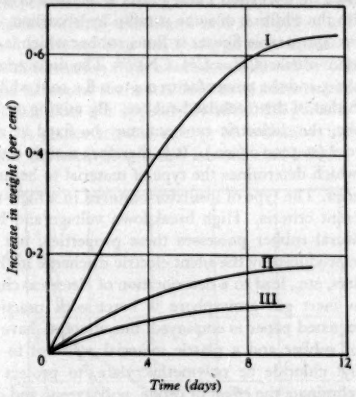


Figure 4. Uptake of water by resins at 20° c. I, Acrylate. II, polyvinyl chloride, 23 per cent plasticizer. III, Polyvinyl chloride 30 per cent plasticizer.

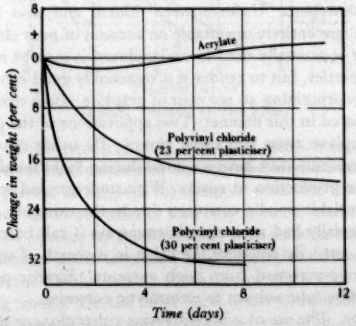


Figure 5. Change in weight of resin on immersion in oil at 90° c.

produce copolymers of these molecules having all the desired properties for any given purpose has yet to be investigated.

Submarine cables. The desiderata of an insulator for use in a submarine cable are now very well defined. The dielectric constant should not exceed 2·5 and the

power factor 3×10^{-3} while water absorption should be at a minimum. In the materials at present in use in this country these figures are obtained by thoroughly deresinating guttapercha and mixing with pure petroleum jelly, giving the so-called *K*-gutta; in America, the so-called Paragutta is a mixture of guttapercha, deproteinized rubber with the addition of some paraffin hydrocarbon. The only synthetic product which gives comparable figures is Buna rubber which is an interpolymer of butadiene with acrylic nitrile (CH_2=CH.CN)[19]. The dielectric constant is said to be between 2·2 and 2·5 and the power factor 0·4 to $0{\cdot}8 \times 10^{-3}$ while water absorption is comparable with that of deproteinized rubber. By mixing synthetic rubber with purified guttapercha, the dielectric constant may be fixed at 2·6 and the power factor at 1 to 2×10^{-3} (at 5000 c./sec.). It is therefore mainly a question of relative costs of products which determines the type of material to be used in such cables.

High-tension cables. The type of insulator required in a high-tension cable must satisfy rather different criteria. High breakdown voltage and high resistivity are essential. Raw natural rubber possesses these properties, but is easily oxidized especially by ozone produced by the silent electric discharge in the cable. Addition of antioxidants, fillers, etc., lead to a deterioration of electrical characteristics, while preservation in an inert gas atmosphere is beset with practical difficulties[20]. Normally oil-impregnated paper is employed, but attempts have been made to use composite layers of rubber and a plastic material resistant to ozone[21], such as plasticized polyvinyl chloride or polymethacrylate, to protect the rubber from disintegration. To eliminate the effect of ozone, polystyrene and chlorinated rubber have been suggested[22], but it seems that complete immunity from attack can only be obtained by using chloroprene, interpolymers of vinyl chloride and vinyl acetate or polyisobutylene. Unfortunately thiokol and also certain butadiene rubbers (Perbunan) are entirely unsuitable on account of poor electrical properties.

Though rubber is normally used as an insulator, it may be necessary to retain its mechanical properties, but to render it a reasonably good conductor[23]. This is accomplished by incorporating 40 per cent of graphite. Both natural and synthetic rubbers may be treated in this manner. Two applications of the product have been made, namely aeroplane tyres and cable covers. By using conducting tyres any charge which the machine may have acquired during flight is safely discharged on landing without the production of sparks. With underground lead-covered cables there is often appreciable anodic corrosion due to the existence of earth currents. The trouble is especially bad near electric tramways if rail bonding has not been thoroughly carried out. By encasing the cable in a sheath of conducting rubber, the lead is effectively screened from such currents. Similar protection may be offered to any metallic tube subject to electrolytic corrosion.

Film permeability. The use of synthetic resins either alone or mixed with natural products for protective coatings on metallic substrates is now firmly established. There is, however, not much published physical data about the behaviour of such films, though it appears to be a general rule that products of lower molecular weight than those used for bulk resins is desirable in order that the films may be sufficiently flexible. A second factor is the permeability to solutions containing ions likely

to attack the protected surface. This permeability may be most conveniently determined by measuring the internal resistance of a cell consisting of an anode upon which is deposited the resin and a platinum cathode separated by the corroding solution[24]. In the experiments of Young and Gerhardt a 0·11 per cent solution of sodium chloride acidified with hydrochloric acid to a *p*H of 4·2 was used. As is evident from figure 6 complete film continuity is only attained under their conditions, that is, by film formation from a solution followed by baking, at quite high film thicknesses. The nature of the metal does have some marked effect, although the nature of the surface does not appreciably affect the porosity.

Reinforced resins. A glance at table 1 on p. 68 will show immediately that the tensile strength of the thermosetting resins is somewhat inferior to that of natural

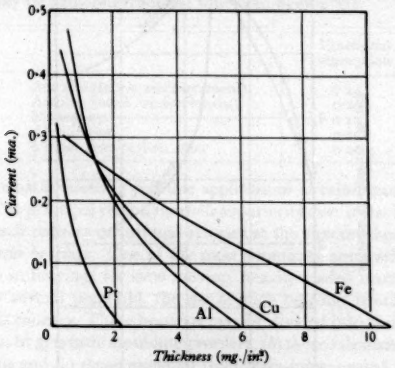


Figure 6. Porosity of resin films.

products. In view of the extensive cross-linking in resins it is quite impossible to orientate the molecules in any specified direction by mechanical means in order to improve their properties, as happens for example when the molecules of cellulose in artificial silk may be so orientated. The only method available is to find what improvement can be effected by reinforcement with suitable materials.

If bakelite is stressed under tension, Young's modulus decreases with increasing stress until the resin breaks down; for a cresol resin the tensile strength is 5000 lb./in.² (3·5 kg./mm.²). On the other hand under compressive stress, the value of the modulus increases (for a cresol resin) from 3×10^5 lb./in.² (210 kg./mm.²) to about 6×10^5 lb./in.² (420 kg./mm.²) at a stress of 6000 lb./in.² (4·2 kg./mm.²), failure now occurring at about 20,000 lb./in.² (14·1 kg./mm.). For the display of the most favourable properties therefore the resin should be in compression. Some improvement may be effected by the use of a wood filler, but unfortunately at a practically corresponding increase in density. Much better results[25] may be obtained by the

use of continuous reinforcement with cord or woven fabric when compressive and tensile strengths are then more nearly equal. It is not essential to have the elastic moduli of resin and reinforcement identical to prevent separation of the components under stress, although it happens that cotton and phenol-formaldehyde resin are similar in these respects. Figure 7 shows the effect of cord reinforcement on the characteristics of the resin in compression [25]. In compression the value of Young's modulus is not much altered up to a stress of 6000 lb./in.^{2} (4·2 kg./mm.^{2}). Thereafter, however, the modulus falls very rapidly in the reinforced material. The effect of reinforcement is very marked when the composite resin is in tension. The tensile strength rises from 5000 to 45,000 lb./in.^{2}* and the modulus of elasticity goes up to

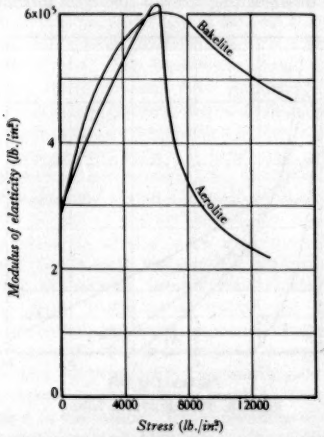


Figure 7. Variation of modulus of elasticity of bakelite and of aerolite under compression.

7×10^6 lb./in.^{2}* Again there is a marked decrease in the latter quantity above a stress of 6000 lb./in.^{2} In point of fact the elastic properties are really determined almost wholly by the cord, for example, the strength of a specimen containing 118 threads was 1150 lb., that of the threads alone was 1190 lb. Moulding pressure has also a profound effect in modifying elastic properties. Too low a pressure yields a brittle material similar to bakelite. The most effective pressure appears to be between 1 and 4 ton/in.^{2} When such a pressure is applied to the liquid resin during impregnation the cord will therefore be in a high degree of tension, which tension will be preserved when the resin cools and hardens. It is significant that the maximum value of the modulus occurs at a stress of 6000 lb./in.^{2} for, assuming a value of Poisson's ratio of 0·3, this corresponds to a lateral tension of 2000 lb./in.^{2}—a

* Private communication from Dr N. A. de Bruyne.

figure close to the moulding pressure. The marked progress made in the above-mentioned connexions shows that when resins can be made to closer molecular specifications, these reinforced materials may find a wide range of applications.

Energy absorption in reinforced resins. In figure 2 on p. 72 the ratio of the area of *OABCO* to *OABD* represents the fraction of energy absorbed by an elastic material during a stress cycle. Such energy will of course be dissipated as heat. Provided the strain is small the energy absorption is independent of strain and remains constant throughout a large number of cycles. Energy absorption is characteristic of amorphous and of polycrystalline materials: it is small in single crystals[26]. The energy must therefore be dissipated in the interstices of the resin, which are numerous (cf. p. 70). Some idea of the high energy absorption of reinforced resins may be gathered from the following figures[25]:

	Fractional absorption
Aerolite (fabric reinforcement)	0·24
Aerolite (cord reinforcement)	0·20
Mahogany	0·12
Aluminium	0·011
0·9 per cent carbon steel	0·0017

One of the most interesting possible applications of reinforced resins is for air-screw blades. The principal reason for their superiority over metal blades lies in their low density which reduces centrifugal stresses at the airscrew boss to a minimum.

Synthetic resin bearings. One of the most promising applications of reinforced resins has been in bearings for slow moving, heavily loaded machinery. Although the idea is now several years old, the use of such bearings is only now becoming extensive in this country. These bearings may be divided into three types, (*a*) those made from plain or graphitic moulding powders, (*b*) those fabricated from laminated reinforced resins and (*c*) those moulded from resin-impregnated felted fibres. The first type has only limited applicability for small lightly loaded bearings, running dry or lubricated, for example, spindles of textile machinery. The second and third types, however, find their greatest application in roller mill bearings. Their superiority over metal bearings is that frictional losses are cut down by as much as a factor of three as is shown by table 3[27].

Table 3. Coefficients of friction of laminated bearing materials and of brass

Lubricant	Brass	Type 2 Synthetic resin laminated (cotton base)	Type 3 Ligno-cellulose base, felted
Dry	—	0·25–0·42	—
Grease	0·10	0·14–0·12	0·1–0·15
Oil	0·06	0·07	0·045–0·075
Emulsion	0·085	0·05–0·08	0·07
Water	0·09–0·2	0·03–0·07	0·02–0·07

Also a very cheap lubricant is employed, namely, water. Moreover, should the bearing seize, no damage is done to the journal; small particles of abrasive are soon embedded in the resin and thus do not score the journal especially if the latter is surface-hardened. The only disadvantage is the low thermal conductivity of the resin—about 1/500 of that of ordinary bearing metals—an ample supply of cooling water coupled with a suitably designed anchorage being necessary to maintain a low coefficient of friction. Without lubrication the bearing acts as a most efficient brake lining. For large bearings it is usual to insert the laminated material as strips in a bent steel plate, thus facilitating water cooling and replacement. On the other hand, it is impracticable to use such laminated materials for small or intricate bearings on account of the difficulty of machining the reinforced resin. Recently, therefore, a moulding material of the thermosetting type has been produced in which a felted mat of cellulose fibres—similar in structure to that formed in paper making—is impregnated with a partially polymerized resin precipitated from

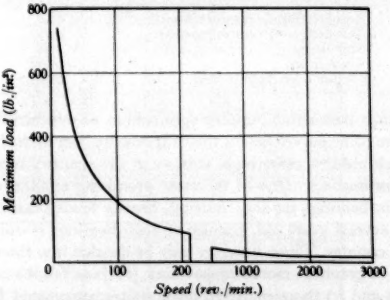


Figure 8. {Load, speed} diagram for graphitized resin. Diameter of bearing 1¼ in.

aqueous solution or by treatment with an emulsion of the resin. The impregnated felt is then injected into the mould and polymerization completed by heat treatment. On account of the random orientation of the reinforcement, the tensile strength drops to 8000 lb./in.² compared with 10,000 lb./in.² for the laminated material. The compressive strength is 20,000 lb./in.² compared with 40,000 to 45,000 lb./in.² at right angles to the laminae and 20,000 to 24,000 lb./in.² parallel to the laminae for materials of type 2. The coefficients of friction are similar for both types. Figure 8[27] indicates the maximum permissible load for a bearing, assuming a maximum temperature of 65° C., for various speeds of a ½-in. shaft, there being no lubrication except that afforded by the graphite incorporated in the bearing. The addition of graphite, however, introduces the troublesome feature of electrolytic corrosion, the graphite forming the negative pole of the cell. This difficulty has been successfully overcome by using very finely divided graphite and by the incorporation of chlorinated hydrocarbons in the resin to decrease water absorption.

The precise reason for the superiority of resin bearings is not yet understood.

Tentatively the following hypothesis has been advanced[27]. First, the surface of the bearing consists of a forest of cellulose fibres, standing proud of the surrounding resin. These absorb water and under load probably form a matted surface thoroughly saturated with water which cannot easily be squeezed out—a process which would increase the coefficient of friction. Secondly on account of the low modulus of elasticity the bearing surface will more readily accommodate itself to any slight irregularity on the shaft of the journal thus eliminating areas of high pressure likely to remove the lubricant and so lead to seizure.

REFERENCES

(1) STAUDINGER. *Trans. Faraday Soc.* **32**, 323 (1936).
(2) HOUWINK. *Trans. Faraday Soc.* **32**, 126 (1936).
(3) DE BOER. *Trans. Faraday Soc.* **32**, 10 (1936).
(4) HOUWINK. *Elasticity, Plasticity and Structure of Matter*, p. 3 (Cambridge, 1937).
(5) DE BOER. *Trans. Faraday Soc.* **32**, 10 (1936).
(6) BRAZIER and HOLLAND-BOWYER. *Lubrication Discussion, Institution of Mechanical Engineers*, **1**, 34 (London, 1937).
(7) HOUWINK. *Elasticity*, p. 160.
(8) WHITBY. *Trans. Instn Rubber Industr.* **5**, 184 (1929); **6**, 40 (1930).
(9) MARTIN and PATRICK. *Industr. Engng Chem.* **28**, 1146 (1936).
(10) HOUWINK. *Elasticity*, chap. 6.
(11) MEYER, VON SUSICH and VALKO. *Kolloidzschr. Z.* **59**, 208 (1932).
MEYER and FERRI. *Helv. Chim. Acta*, **18**, 570 (1935).
GUTH and MARK. *Mh. Chem.* **65**, 93 (1934).
KUHN. *Kolloidzschr. Z.* **68**, 2 (1934); **76**, 258 (1936).
MARK. *Proc. Rubber Technology Conference* (London, 1938).
(12) CURME and DOUGLAS. *Industr. Engng Chem.* **28**, 1123 (1936).
(13) STAUDINGER. *Die Hochmolekularen Verbingdungen* (Berlin, 1932).
(14) SIGNER and GROSS. *Helv. Chim. Acta*, **17**, 59, 335, 726 (1934).
(15) See also STAUDINGER. *Trans. Faraday Soc.* **32**, 11 (1936).
(16) BLAIKIE and CROZIER. *Industr. Engng Chem.* **28**, 1155 (1936).
(17) NAUNTON and WARING. *Proc. Rubber Technology Conference* (London, 1938).
(18) NOWAK and HOFMEIER. *Zur Entwicklung der Chemie der Hochpolymeren*, p. 104, Verlag Chemie (Berlin, 1937).
(19) BADUM. *Proc. Rubber Technology Conference* (London, 1938).
(20) British Patent, 456,340, 456,659, 461,371, 465,321 (1936).
(21) Austrian Patent, 151,219 (1936).
(22) German Patent, 652,749 (1937).
(23) Compagnie générale d'Electricité, French Patent, 782,956, 804,684; British Patent, 447,064, 455,193 (1936).
(24) YOUNG and GERHARDT. *Industr. Engng Chem.* **29**, 1277 (1937).
(25) DE BRUYNE. *J. R. Aero. Soc.* **41**, 523 (1937).
(26) See BURGER'S section on HOUWINK'S book on *Elasticity*.
(27) EYSSEN. *Lubrication Discussion, Institution of Mechanical Engineers*, **1**, 84 (London, 1937). Cf. also RICHARDSON, ibid. p. 249; ROCHESTER, ibid. p. 260; WATSON-SMYTH, ibid. p. 277; BEUERLEIN, ibid. p. 8.
HORNE. *Iron Steel Engr*, **12**, 12 (1935); *Engineer*, **162**, 270 (1936).

The following are of general interest in connexion with plastics:

(*a*) Faraday Society. *Discussion on Polymerization*, **32**, 1–412 (1936).
(*b*) ELLIS. *Synthetic Resins*. Reinhold Publishing Corp. New York.
(*c*) HOUWINK. *Elasticity, Plasticity and Structure of Matter* (Cambridge, 1937).
(*d*) *Proceedings of the Rubber Technology Conference* (London, 1938).
(*e*) *Lubrication Discussion of the Institution of Mechanical Engineers* (London, 1937).
(*f*) *Zur Entwicklung der Chemie der Hochpolymeren*, Verlag Chemie (Berlin, 1937).
(*g*) *Symposium on Plastics, Industrial and Engineering Chemistry*, **28**, 1144 (1936).

RECENT PROGRESS IN ASTRONOMY

By A. C. D. CROMMELIN

An account of recent astronomical progress naturally begins with the Sun; but progress in solar physics has been mainly gained by the spectroscope, and is described in another section; one of its results may be referred to here, namely that the sunspots, which are now known to be of the nature of vortices, reverse their direction of spinning at the beginning of a new sunspot cycle, so that the period after which conditions are completely restored is twice the period of 11 years in which the spots complete their cycle of waxing and waning.

Use has been made of the very strong gravitational field round the Sun to test Einstein's prediction that rays of light would suffer bending on traversing a strong gravitational field. The field round the Sun is the only one within our reach that is sufficiently strong for the test. At most of the total eclipses of the Sun from 1918 onwards photographs have been taken of the stars round the eclipsed Sun. Comparison of these with others, taken when the Sun had moved away, has closely verified the predicted shift, which is $1\frac{3}{4}''$ at the limb of the Sun and varies inversely as the distance from its centre. This is just twice the shift that a material particle, travelling with the speed of light, would undergo on passing the Sun's limb under Newton's law.

There has been a great increase in the estimates of the past duration of the Sun as a distributor of light and heat. Half a century ago Lord Kelvin assigned an upper limit of 20 million years, on the assumption that the Sun's heat was derived solely from the coming together of its particles from a great distance. The discovery of the energy that is latent in the atom increased the possible duration to millions of millions of years, which was more than enough to satisfy the geologists, whose demand is limited to about 3000 millions of years.

There has been much controversy over the manner of origin of the planetary system. Laplace's nebular hypothesis held its own for most of the last century, but has now been abandoned "not because it is wrong in theory, but because it fails in practice"(1). As an alternative, various theories of tidal action on the Sun from the near approach of another star, resulting in the expulsion of the planets, were proposed; it is now generally admitted that these break down through failure to explain the large "moment of momentum" of the planets, as compared with that of the Sun; the term denotes the speed of each planet multiplied by its mass and by its distance from the Sun.

R. A. Lyttleton has recently proposed a new theory to overcome this difficulty. He supposes the Sun to have formerly had a companion Sun revolving about it, at about the distance of Uranus; an external star made such a near approach to this companion that it enticed it away from its allegiance to the Sun; while the two

were in proximity, a filament of matter is supposed to have formed between them by tidal action; it was claimed that part of this filament might have remained under the gravitational control of the Sun, and formed the matrix of the planets. This theory is still under discussion, but several difficulties have been pointed out, and the origin of the planetary system must still be considered an unsolved mystery.

Leaving their origin in abeyance, I proceed to describe recent study of the present condition of the planets. At the partial transit of Mercury on 11 May 1937, careful search was made for a ring of light surrounding the portion of the planet outside the Sun, but none was seen. Venus, when commencing to transit, shows a conspicuous ring round the part outside the Sun, due to refraction of sunlight by the planet's atmosphere; it is concluded that Mercury has little or no air, and that the temperature of its sunward face must approach that of molten lead. It is now generally agreed that it rotates in the same period as its revolution, 88 days. This was first announced by Schiaparelli and Lowell in the last century, but has been slow in obtaining general acceptance. Owing to the high eccentricity of the orbit and the consequent variation in its orbital motion, $\frac{3}{8}$ of Mercury's surface has perpetual sunshine, $\frac{3}{8}$ has perpetual night, and $\frac{1}{4}$ alternates between day and night. Mercury played an important part in the verification of Einstein's theory of gravitation. Its perihelion point had long been known to advance by an amount exceeding that calculated on Newtonian theory by 43″ per century. This agreed exactly with Einstein's predicted excess, which was based entirely on theory.

This undesigned agreement drew general attention to Einstein's theory. His other predictions were that light would be bent in a gravitational field to twice the amount of a particle moving with the speed of light under the Newtonian law; and that such a field would deflect lines in the spectrum towards the red. The first prediction was verified by photography of stars round the eclipsed Sun in 1919 and 1922; the other by photographs obtained by Adams at Mt Wilson of the spectrum of the companion of Sirius, which has an extremely high density. It was a veritable intellectual triumph to predict these effects by pure reasoning, in advance of their detection by observation. It was also a notable photographic triumph to photograph the spectrum of the companion so close to the overpowering radiance of Sirius.

There has been little increase in knowledge about Venus; this planet is permanently cloud-covered, and its period of rotation remains unknown; but the spectroscope failed to show such a shift of lines as would correspond with a period of rotation near 24 hr.; the period seems to be several days, but probably much shorter than 225 days, the period of revolution round the Sun; in that case the night side of the planet would be much colder than the day side, which is not verified by observations at Mt Wilson. These observations also failed to show the lines of oxygen and water-vapour; but owing to the thick veil of clouds only the high regions of the atmosphere of Venus could be studied; these gases may be present at lower levels. There has been no great increase in knowledge regarding the

surface of the Moon; but continued observation brings to light many fresh small details on the surface; thus H. G. Tomkins traced some long dark markings, which he concluded to be ancient streams of lava; W. H. Pickering noted changes of tone and of tint in certain regions, the cycle of change being the same as the lunation, 29½ days; he ascribed these to some form of vegetation, and was doubtless right in stating that the period of such changes on the moon would be 29½ days; the moon's axis is almost perpendicular to the ecliptic, so there are no appreciable annual changes of temperature; but the apparent absence of air and water on the Moon makes the vegetation hypothesis difficult to accept.

Careful observations of the temperature of the lunar surface, and of its changes during a lunar eclipse, made by Pettit and Nicholson in 1927, showed that the absolute temperature of the point examined fell from 342° at the beginning of eclipse to 156° during totality, rising to nearly the original value at the end of eclipse. H. N. Russell concluded from these measures that the surface material of the Moon is a very poor conductor of heat; he suggests that it is like pumice or volcanic ash, rather than solid rock; he says "While the surface is at times hotter than boiling water, at other times almost as cold as liquid air, the temperature a foot below probably varies but a few degrees from an average near the freezing point"[2].

Since the angular motion of the Moon greatly exceeds that of other bodies, it is the best object to choose for detecting changes in the Earth's rate of rotation. Halley was the first to detect an apparent acceleration of the Moon, by comparing records of eclipses given by Ptolemy, and later by Arabian observers, with those made in his own time. The acceleration was first explained by Laplace as due to a slow change in the eccentricity of the Earth's orbit. Adams detected an error in Laplace's work; the increase in longitude in a century should be 7″, not 11″; as the observed increase is 11″ there must be some other cause in addition to the one named by Laplace. The accepted explanation is an actual increase in the length of the day, due to the friction of the tides; the Sun and the inner planets give evidence in favour of this, as they too show small accelerations. Drs Jeffreys and Taylor have discussed the tidal friction in the Irish Sea, and in several other enclosed Seas, above all the Bering Sea; their results, while not claiming strict accuracy, show that tidal friction is amply sufficient to explain the excess of acceleration over that due to the Laplace term. The latter term arising from planetary perturbation of the Earth's orbit will in the course of ages reverse its sign, but the tidal action will always be in the direction of lengthening the day; it is, however, accompanied by a slow recession of the Moon from the Earth, which will diminish the amount of apparent acceleration. If the system lasts long enough, tidal friction will at length cause the Earth to turn one face constantly to the Moon, the day being 47 times its present length; this will happen in 5×10^{10} years[3].

A great improvement in the predicted positions of the Moon was brought about in 1923, when Prof. Ernest W. Brown's Tables replaced those of Hansen, which had been in use for more than half a century. The new tables are practically perfect as regards all terms of short period, but there are still some unexplained fluctuations

of long period. As the other rapidly moving bodies (Mercury, Venus, and Jupiter's satellites) appear to show similar fluctuations, though smaller in proportion to their rates of motion, it is suspected that they arise from slight fluctuations in the Earth's rate of rotation, possibly produced by some rearrangement of matter in the Earth's interior. Till their cause is understood it is impossible to make strictly accurate predictions of the Moon's position for many years in advance.

There has been some advance in knowledge about conditions on Mars; Dr Wright took many photographs in light of different colours, and found that the images in ultra-violet light were considerably larger than those in infra-red; but the former showed practically no surface-details, while the latter brought out the dusky regions very distinctly; it is concluded that the former photographs show the atmosphere to a height of 50 or 60 miles, while the latter show the planet's surface. Traces of oxygen and water-vapour are found in the spectrum; the amount of oxygen is estimated as 15 per cent of that on Earth, and of water-vapour 5 per cent. It is generally agreed that the orange regions on the planet are sandy deserts, and that the darker regions may be covered with some form of vegetation; their changes of tint in the course of the Martian year support this interpretation; its likelihood is strengthened by radiometric measures of the temperature of the surface; it appears that this may reach 50° Fahr. about midday in the equatorial regions. The nights must be cold, but since there is frequently a whitish border to the disc, indicating cloud or mist at morning and evening, it is possible that night clouds may arrest the escape of heat. Owing to the thinner air the Martian temperature should be more sensitive to changes in the solar radiation than that of the Earth; E. M. Antoniadi announced a correlation between the sunspot cycle and the rate of melting of the Martian polar caps; the rate was greatest at sunspot maximum. The suggestion of Percival Lowell that the Martian canals are irrigation works designed by intelligent beings has failed to secure general acceptance, but it has many adherents, including most of those who have observed the planet from Flagstaff, which is 7000 ft. high, and has clear and steady air.

The discovery of new minor planets goes on continuously; 1420 of them have been given permanent numbers, which are only assigned to those for which reliable orbits have been calculated; some hundreds more have been observed, but not sufficiently to get good orbits. Up to 1898 the zone of these planets was looked on as bounded by the orbits of Mars and Jupiter; these limits are now known to be far exceeded in both directions; in 1898 Eros was found, which comes well inside the orbit of Mars and approached the Earth within about 16 million miles early in 1931; numerous visual and photographic observations were then made to find its parallax, and hence the distance of the Sun; this involves such extensive measures and computations that the results are not yet available. Eros appears to be shaped somewhat like a dumb-bell; the period of rotation is about 5 hr., and the light varies by 1½ magnitudes when the Earth is in the equatorial plane of Eros. Between 1932 and 1937 four tiny planets were found that can approach still closer to the Earth than Eros; in 1932 Amor and Apollo were found, their least distances being 10 million and 6 million miles; early in 1936 Adonis approached within 1½ million,

and in 1937 Hermes approached within ½ million miles; the closest known approach of any body other than the Moon.

Amor has the peculiarity that it may be three times in opposition at a single apparition. Its velocity at perihelion exceeds that of the Earth, but only for a short time; thus the Earth overtakes Amor, then the latter goes in front again, and finally the Earth again overtakes it. Amor may be used to determine the Sun's distance, as it can approach nearer than Eros; three of its periods make about 8 years, so it will be again near the Earth early in 1940.

Hermes was found by Dr Reinmuth at Heidelberg in October 1937, but passed out of sight after a day or two, but fortunately prediscovery images were found on plates taken at Harvard, Sonneberg and Johannesburg; its least distance from the Earth's orbit is about 220,000 miles, less than that of the Moon.

These Earth-grazing planets are very small bodies, a mile or two in diameter, and can only be seen when very near the Earth; their orbits cannot be found very exactly unless they are observed at a second apparition.

When the first four minor planets were found it was thought that they might be fragments of an exploded planet. This idea has long been abandoned, but a modified form of it has been revived. Prof. Hirayama found, on studying the orbits, that there were five groups of planets, the components of each group having closely similar orbits: there were 133 planets in the five groups; it is conjectured that each group had a common origin, either from the splitting of a single orb, or the separation of a single condensation in the primeval stream of particles. The orbit elements of the whole collection of minor planets show clear traces of the influence of Jupiter: there are gaps, or at least thinner spaces, at distances from the Sun at which the period bears a simple ratio to the period of Jupiter: the most notable gap is where the period is half that of Jupiter; and the richest region is a little inside this gap; there is an analogy to this in Saturn's Ring; the Cassini division corresponds with a period half that of Mimas, Saturn's nearest satellite, and the brightest region of the Ring is a little on the inside of this division. Since the Ring consists of separate particles, the conditions are similar to those in the Minor Planet Zone, and obviously arise from similar causes.

The perihelia of the orbits of the Minor Planets also give evidence of the strong influence of Jupiter; they are packed three times as densely in the direction of Jupiter's perihelion as in the opposite direction; no explanation of the reason of this grouping has been published, but it must obviously be an effect of perturbation.

On the outer border of the Minor Planet Zone lies the so-called Trojan group; its members have the same period of revolution as Jupiter, 11·86 years. They move in accord with the theoretical prediction of Lagrange (1772) that three bodies at the vertices of an equilateral triangle, if projected with velocities proportional to their distance from their centre of gravity, in directions symmetrical with regard to that centre, will continue to form an equilateral triangle; but its size will vary, unless the motion is circular. The conditions of Lagrange are not exactly realized, for the Trojan planets have different inclinations and eccentricities from those of Jupiter; these differences cause oscillations about the equilateral figure, but are

not large enough to destroy the stability of the motion. Twelve Trojans are now (1938) known; seven have longitude about 60° greater than that of Jupiter; the other five about 60° less; the first group is named after Greek heroes, the other after Trojan ones (except Hector in the first group, Patroclus in the second, which were named before the rule was made). The motion of the Trojans is stable as regards perturbations by Jupiter, but it is just possible for Saturn to disturb their configuration and destroy their Trojan character. There is one planet, Hidalgo, which might be an escaped Trojan; its period is 13·67 years; its orbit is so eccentric that its perihelion is near the orbit of Mars, and its aphelion near that of Saturn; it has also the high inclination of 43½°; it is visible only when near perihelion, but it has now been observed at this point a second time after completing a revolution, so that its orbit is quite well known.

The search for Minor Planets is now made entirely by photography; the planet leaves a short trail, while the stars are points; but for very faint planets, the camera is made to follow the estimated motion of the planet, so as to concentrate its light in a smaller space while the stars trail. Some hundreds of new planets are announced every year, but no planet is officially numbered until five observations, extending over several weeks, are available. Among the Observatories undertaking this work, Bergedorf, Heidelberg, Johannesburg, Nice, Uccle, deserve special mention; the Berlin Rechen-Institut undertakes the publication of orbits and ephemerides.

Passing to the four Giant Planets, there have been notable changes of view as to their physical condition. In the last century Lord Kelvin's estimate of 20 million years as the past duration of the Sun was widely accepted, and the conclusion, drawn by R. A. Proctor and others, that the Giant Planets were still very hot, seemed probable. The discovery of the energy latent in the atom led to the discarding of Kelvin's estimate; moreover, measures of the disintegration of uranium and thorium led the geologists to extend the estimated past duration of the Earth to some 3000 million years. Dr Harold Jeffreys then put forward the conclusion that the outer layers of the Giant Planets must be cold; the observers were somewhat unwilling to accept this; however, observations by Coblentz with a radiometer gave for the surface of Jupiter a temperature of −130° C. Many astronomers took this to mean that the Giant Planets had parted with all the original heat, and were cold down to their centres; Dr Jeffreys's suggestion of a thick layer of ice in the interior strengthened their view; he has since explained that he meant hot ice, kept solid by the great pressure; but he has now withdrawn the ice suggestion and says that "Temperature of the order of 100,000° (C.) at the centres (of the Giant Planets) are to be expected".[4] The high temperature of the interiors helps to explain the low mean densities of the Giant Planets, which are about the same as that of the Sun, viz. 1·4 times that of water.

Careful study of Jupiter's surface has given interesting results. The Great Red Spot has been under observation for about 70 years, and is probably connected with objects observed still earlier, by Hooke in 1664 and Schwabe in 1831; a puzzling feature is its variable period of rotation, which has caused it to drift through some 180° in longitude as compared with other objects in the same latitude;

in recent years a marking called the South Tropical Disturbance has been in the same latitude, but having a more rapid rotation it has several times passed the Spot; on these occasions it pushes the Spot with it for a time, but the Spot appears to drift back again to its normal position. It has long been known that the equatorial zone of Jupiter rotates about 5 min. quicker than the temperate zones; recent measures show that on the average the south temperate zone is about 19 sec. quicker than the north one; there are many abnormal movements among the spots; Rev. T. E. R. Phillips and other British observers noted a group of spots that appeared to have a cyclonic motion in a long oval; the meteorology of the planet is very complicated; there are many different atmospheric currents.

Prof. de Sitter made a new determination of the masses and densities of the four great satellites compared with that of the Moon; their masses are 0·985, 0·643, 2·129, 1·174; the densities are 1·06, 1·00, 0·61, 0·38; the low density of the last is surprising. Five small satellites are now known, besides the old four. The fifth was discovered by Barnard in 1892; it is very near the planet and revolves in less than 12 hr., and its perijove makes two revolutions in a year, this being a result of Jupiter's great oblateness; the sixth and seventh are both about 7 million miles from Jupiter, and revolve in 8½ months; the eighth and ninth, discovered respectively by Melotte in 1908 and Nicholson in 1914, are 14½ million miles from Jupiter, and revolve in about 2 years, but in the opposite direction to the others; this peculiarity is shared by Phoebe, the ninth satellite of Saturn; it has been shown to give greater stability for distant satellites. It will be seen that the four distant satellites form two pairs; but, in spite of the equality of their periods, their planes of motion, and the directions of the major axes of their orbits are quite different.

Passing to Saturn, the great difference of rotation period between the equator and higher latitudes has been confirmed; the periods range from $10^h\ 15^m$ near the equator to $10^h\ 38^m$ for the higher latitudes, a range four times as great as that in Jupiter. Occasionally white spots break out with great suddenness on Saturn; the most striking of these was discovered in 1933 by Mr W. T. Hay; it evidently arose from some convulsion far below the surface; this was shown by the rapid lengthening of the spot to the west, indicating that it came from a region of lower rotational speed.

The transparency of the outer ring was proved by the observation in 1917 by Capt. Ainslie and Mr Knight of the passage of a star of the seventh magnitude behind it; the star remained visible, and twice brightened up, indicating regions where the particles forming the ring are less densely packed.

The very distant satellite Phoebe, discovered by W. H. Pickering in 1898, has retrograde motion, and the very long period of 1½ years; its orbit plane is inclined only 5½° to that of Saturn, much less than the other satellites.

The rotation period of Uranus has been determined in two different ways: Lowell in 1911 found by the spectroscope the period $10^h\ 45^m$; and Campbell in 1917 found a periodic variation in its light in $10^h\ 49^m$; this variation must have been due to some prominent markings on the disc; it is to be noted that the planes of the orbits of the satellites are practically coincident, and the plane of the planet's

equator must be the same, since otherwise the attraction of the equatorial protuberance would have different effects on the satellites, and would cause their planes to be different. The value $10^h\ 49^m$ has more weight, as it resulted from observations of the light-variation extending over several days.

No definite markings have been observed on the disc of Neptune, yet the position of its equator and its period of rotation are now known with considerable accuracy; the former was obtained from the shift of the plane of its satellite Triton; the pole of that plane describes a circle round the pole of Neptune in 580 years, and enough of that circle has now been swept out to determine its centre, namely, Neptune's pole. From spectroscopic observations made by Moore and Menzel and from periodic variation in its light, the period of Neptune's rotation is deduced to be about $15^h\ 40^m$, which is considerably longer than that of the other Giant Planets. The direction of Neptune's rotation was found by the spectroscope to be *Direct*, i.e. the same as the Earth's and opposite to the motion of Triton. It is the only case where a satellite near its primary revolves in a direction opposite to the rotation of the latter.

The spectra of the four Giant Planets all have strong dark bands, which have now been identified as due to methane; they get stronger for the more distant planets, presumably owing to lower temperature. Bands due to ammonia are also found in the spectra of Jupiter and Saturn; Dr Dunham explains their absence from the others by the lower temperature there having caused the ammonia vapour to condense. As in the case of Jupiter, a cold exterior is compatible with a good deal of heat inside, though this doubtless is less for the smaller and more distant orbs; their original temperature was presumably lower, and the distance from the centre to the surface was smaller, facilitating the escape of heat.

It was natural that after the discovery of Neptune speculation arose about possible planets outside it. It was found that Lalande had observed Neptune twice in 1795; his observations helped to determine Neptune's orbit; but it was considered that the well-determined orbit of Uranus gave more hope of finding indications of external disturbances. The two researches that merit record, since they led to the discovery of Pluto, are those of Percival Lowell and W. H. Pickering; the former was published in 1915, shortly before Lowell's death; it was based entirely on the observations of Uranus, which then extended over nearly three revolutions, though the earlier observations, made before Herschel's discovery in 1781, were few and rough; he obtained an orbit which is extremely close to the actual orbit of Pluto, as the following table shows:

	Lowell	Actual		Lowell	Actual
Mean distance from Sun	43·0	39·5	Ascending node	Not given	109°4
Period (years)	282	248	Inclination	10°	17°1
Longitude of perihelion	204°9	223°4	Longitude 1930·0	102°7	108°5
Date of perihelion passage	1991·2	1989·8	Magnitude	12 to 13	15
Eccentricity	0·202	0·248	Mass (Earth = 1)	6·7	?

The agreement is closer than expected, considering the small residuals between observation and calculation, on which the research depended; but the observed magnitude was considerably fainter than the predicted one, so that the predicted mass is probably too high; this would imply that Lowell overestimated the height of the perturbation waves, while locating their positions with fair accuracy; it will be seen that the predicted position at discovery in 1930 was only 6° wrong. W. H. Pickering gave three different predictions; the first, made in 1909, was rough, and I only give the predicted longitude in 1930, which was 135°; the other two appeared in 1919 and 1928; they are given below:

	Pickering		Actual		Pickering		Actual
	1919	1928			1919	1928	
Mean distance from Sun	55·1	30·1	39·5	Ascending node	100°	180°	109°4
Period (years)	409·1	164·8	248	Inclination	15°	Not given	17°1
Longitude of perihelion	280°1	252°5	223°4	Longitude 1930·0	102°6	135°	108°5
Date of perihelion passage	1720·0	1973·8	1989·8	Magnitude	15	13	15
Eccentricity	0·310	0·195	0·248	Mass (Earth = 1)	2	0·75	?

On the whole the 1919 prediction is better than the 1928 one; it gives a close approximation to the node and inclination, also to the magnitude; and its value of the mass is probably the best of the forecasts. In spite of the resemblance between some of the forecasts and the actual elements of Pluto, the majority of astronomers have followed Prof. E. W. Brown's suggestion that the predictions were illusory, and the agreement of Pluto with them a mere fluke. The present writer rejects this suggestion, but leaves readers to draw their own conclusions. The accurate observation of Uranus and Neptune will ultimately give a reliable mass for Pluto. Its orbit elements are already well determined, as many images on plates exposed some years before the discovery, the earliest being in 1914, were found. At perihelion, which will occur in 1989, Pluto comes slightly inside Neptune's orbit; but since three periods of Neptune are almost exactly equal to two of Pluto, we have to go thousands of years backward or forward to find a conjunction of Neptune and Pluto occurring near Pluto's perihelion. Pluto will be in conjunction with Uranus in 1966, at a distance of 15 units.

Comets. Some more periodic comets with short or moderate periods have been found; the most interesting is Comet Schwassmann-Wachmann (1), discovered at Bergedorf in 1927, but having passed perihelion on 15 May 1925; it has a period of 16·3 years, its orbit lying wholly between those of Jupiter and Saturn; it is the first known comet that is observable round the whole orbit; it has been seen almost every year since 1927. It is subject to occasional outbursts of light, during which it shines with a hundred times its normal lustre; it would seem that these outbursts arise from some explosive action, similar to that noticed in Holmes's comet in 1892. Another interesting comet was found by Forbes in South Africa in 1928; it

proved to be identical with a comet seen by Pons in 1818, and with one seen by Coggia and Winnecke in 1873; its period is about 28 years, its orbit reaching nearly to that of Uranus. It is almost certainly identical with a comet seen by Toscanelli in 1457, and possibly with one recorded by Kepler in 1625.

Encke's comet last returned to perihelion in December 1937; it has not been missed at a single return since 1818; it reached naked-eye visibility in 1937, and does not seem to have lost much light since its first detection in 1786. Its unbroken record is due to its near approach to the Sun, inside the orbit of Mercury, when its strong illumination by the Sun makes it a conspicuous object. The gradual shortening of its period, which was evident from its discovery till 1838, has now practically ceased.

The Comet Pons-Winnecke approached the Earth within $3\frac{1}{2}$ million miles in 1927; it had a star-like nucleus about a mile in diameter. This comet comes very near Jupiter every 12 years; these approaches have caused a notable recession of its perihelion point, which is now 9 million miles outside the Earth's orbit. On one occasion only, on 28 June 1916, the meteors associated with this comet met the Earth. Another interesting short-period comet was found by Skjellerup in South Africa in 1922; Dr Merton proved its identity with one seen by Grigg in 1902; its period is almost exactly 5 years, and it has been seen at every return since 1932; it is liable to very large perturbations by Jupiter, and appreciable ones by the Earth. Another comet that can approach very near Jupiter was found by Gale in 1927, and returned in 1938; its period is 11 years, and it evidently approached Jupiter very closely in 1917, when its orbit must have suffered change. It has now been observed on its return to perihelion 18 June 1938.

Photography of the tails of bright comets has frequently shown luminous knots receding from the head with accelerated motion, indicating a repulsive force much stronger than gravitation; presumably electrical repulsion.

It has been established by Prof. Strömgren and his assistants that several comets that passed the Sun with hyperbolic motion were moving in long ellipses when they entered the planetary system, the change to hyperbolic motion arising from planetary perturbations. There is thus no evidence that any comet has entered the Solar System from the stellar regions, and it is concluded that they all had their origin within the Solar System, or at least in outlying matter connected with it, and sharing its motion through space. Further than this we cannot go till the mystery of the origin of the planetary system is solved; hitherto all the attempts made in this direction have been found to involve some mathematical difficulty. However meteors do not in all cases resemble comets in this respect. Harvard College Observatory recently sent an expedition to Arizona which made photographic observations of meteors from two stations a few miles apart, thus determining their height and velocity; the latter exceeded the parabolic value in a large number of cases, indicating that these meteors were visitors from the stellar regions.

The Stars. A great improvement has been effected in the technique of determining stellar parallaxes. Photographs are taken about 6 months apart, on or near

the meridian; reliance is chiefly placed in observations of right ascension, which are unaffected by differential refraction; the probable errors of the measured parallaxes have been reduced to $0''.01$ or less. But this direct method of determining distances is available over only a very small portion of our immense star-system. Of the indirect methods used for more distant regions, that of proper motions was the first used. All the stars down to the ninth magnitude have now been observed with meridian instruments over a period of half a century or more. The great astrographic chart of the whole sky, in which stars down to the twelfth magnitude are being measured, has led to the determination of many new proper motions. The largest known motion, $10''.3$ per annum, belongs to the Barnard star in Ophiuchus, magnitude 9·7; our nearest known neighbour, named Proxima by Dr Innes, is a distant companion of Alpha Centauri, round which it is supposed to revolve in about a million years; its magnitude is $10\frac{1}{2}$. The parsec was introduced by Prof. Turner as a convenient unit for measuring stellar distances; it is the distance, about $3\frac{1}{4}$ light-years, corresponding to a parallax of $1''$; the term "absolute magnitude" was introduced at about the same time; it is the magnitude which a star would have if placed at a distance of 10 parsecs from us. It is found that stars of the same spectral type, but of different absolute magnitude, show slight differences of intensity in certain lines of their spectra; hence study of these lines gives an idea of the star's absolute magnitude; from comparison with the apparent magnitude the distance can be deduced. The proper motions of stars give a rough clue to their distances; obviously on the average the quick movers are the nearest; but the method cannot be pressed to individual cases, owing to the different inclination of the motions to the line of sight, and the different real velocities of the stars.

A powerful spectroscopic method of determining the distances of certain distant stars was found by Prof. Plaskett. The orbital motion of binary stars causes variation in their rates of approach or recession, and a consequent shift in their spectral lines. Now certain lines due to calcium and sodium were found not to share in this shift; examination showed that the gases causing these lines were not in the stars themselves, but in interstellar space, where they appear to be distributed uniformly, or nearly so; thus the strength of the lines in a spectrum gives a good measure of the star's distance.

Another very powerful method for distant objects was announced by Miss Leavitt; she examined a series of photographs of the smaller Magellanic Cloud, and found several variables of the type known as Cepheids (from Delta Cephei, the first of the type to have its light-changes studied). She found that the brighter Cepheids had the longest periods, and that the period of light-change gives a good measure of the absolute magnitude of a Cepheid. A discussion of the nearer Cepheids gave the scale of distances corresponding to different magnitudes; this was not easy, as they are all too distant to obtain reliable distances by the method of parallax; a careful discussion of proper motions and velocities in the line of sight, conducted by Prof. Shapley, gave the solution; since the brighter Cepheids are visible at immense distances, the method is very valuable, and has led to important results both within and beyond our own star-system.

It has been recognized since the days of Sir William Herschel that the Galaxy is the more distant portion of the great star-system to which the Sun belongs; it has now been found possible to locate the centre of this system, which lies in the star-clouds of Sagittarius. Star-density, which is measured by the number of stars per square degree, increases steadily as the Galaxy is approached; but there are irregularities due to extensive clouds of obscuring matter. The most conspicuous of these clouds are the two "Coal Sacks" in rich regions of the Galaxy; one of them is near the Southern Cross. It is thought that similar clouds veil the central regions of the Galaxy, and also account for the splitting of the Galaxy into two streams through the veiling of the middle regions. Now that the kinship of the Galaxy to the Spiral Nebulae is established, it is considered that the Galaxy may be a spiral, but it cannot be seen as such by observers in its plane.

It has long been recognized that there must be some sort of rotation in the Galaxy, to prevent its collapse; it is only in recent years that the details of this rotation have been established; earlier investigators were seeking signs of it among our nearer stellar neighbours; but these near neighbours have practically the same motion as the Sun; we must examine the motion of distant objects, using their velocity in the line of sight as determined by the spectroscope. Dr Oort of Leiden was the first to indicate the sort of effect to look for. Assuming the motion to be one of rotation about the centre, with speed depending on the distance from the centre, then stars in the direction of the centre or the opposite direction will show no radial motion; stars at right angles to these directions and at moderate distances from us will also show none, for their motion is similar to our own; the maxima of speed will be in directions of 45° or 135° from the centre; the motion being one of approach in two of these directions, and of recession in the other two. The spectroscopic observations of distant stars (types B and O being selected as being the most luminous and thus visible at greater distances) were undertaken by Dr J. H. Plaskett with the 72-in. reflector at Victoria; his results, combined with others, led to the conclusion that the whole assemblage of stars forming the Galaxy is rotating in a period which varies for different regions but is about 250 million years for the solar regions, about a centre in Sagittarius, near the brilliant star-clouds in that region, but the actual centre is thought to be hidden by obscuring clouds. Assuming the above period, the Sun is 11,000 parsecs from the centre, and the total mass of our star-system is 150 thousand million Suns. Galactic rotation explains the puzzling fact that stars of high velocity (relatively to the Sun) were all moving more or less in the same direction. When corrected for galactic rotation these stars are "slow movers". Quick movers in the opposite direction would have escaped from the Galaxy.

Two stars will be described in some detail as examples of Giants and Dwarfs; the first is a sixth magnitude star in Monoceros, known as the Plaskett star; it proved to be a spectroscopic binary, the masses of the components being at least 76 and 64 times that of the Sun; their centres are about 60 million miles apart, the stars revolving round each other in a fortnight; their diameters are 20 and 18 times the Sun's, their luminosities 15,000 and 12,000 times the Sun's; their distance is about 9300 light-years. At the other extreme is a small star about 30′ south of

υ Draconis, R.A. (1900) $19^h\ 1^m\ 8^s$, N. Decl. 70° 30′; it was found by the photographs taken for the Greenwich Astrographic Catalogue to have an annual proper motion of $+0''.105$ in R.A., $+0''.508$ in Decl.; its spectrum indicated that it is of type O, the hottest of all the types; its parallax is being redetermined, but it is known to be about $0''.065$; its magnitude is 12·7; it is deduced that the star is smaller than the Earth, but more massive than the Sun, so that a cubic inch of it would weigh about 600 tons. The first white dwarf to be studied was the companion of Sirius; this faint star was discovered by Alvan Clark in 1862; its mass was found to be about half that of Sirius, and it was presumed to have a dull red surface to explain its faintness. But in 1915 Adams at Mt Wilson succeeded in photographing its spectrum, and found that it is a white star like Sirius, but from its faintness its diameter is inferred to be only about three times that of the Earth, so that a cubic inch of it weighs about a ton. The white dwarfs are intrinsically very faint, and can only be seen when they are comparatively near the Sun; so that, although only a few are known, they are likely to form an appreciable fraction of the stars as a whole, and to illustrate an alternative final stage of stellar life; the other alternative is the deep red M type to which the Barnard star in Ophiuchus probably belongs. It has been suggested, with considerable probability, that the central stars of planetary nebulae are white dwarfs. A relation between the mass of a star and its total output of light or absolute magnitude was deduced by Prof. Eddington from theoretical reasoning, and confirmed by testing on stars of known mass and luminosity; previously the mass could be found only in the case of double stars; it can now be found by Eddington's formula for any star of known distance and magnitude. It is applicable to the red dwarfs, but not to the white dwarfs, which have too low a luminosity in proportion to their mass. The smallest stellar masses found are about one-eighth of that of the Sun; Eddington concluded that smaller masses would not attain sufficient temperature to shine as suns.

In 1918 Prof. Shapley published an important research on the Globular Clusters. He used Miss Leavitt's result that the absolute magnitudes of Cepheid Variables could be deduced from their periods of light-variation. The clusters proved to be far more distant than had been previously supposed, and their brighter members are giant stars. The centre of the system of clusters is found to be the same as the centre of galactic rotation; the clusters form two groups, one on each side of the galactic plane; but as the system of clusters is smaller than the Galaxy, there are no clusters seen in the direction opposite to the galactic centre.

The principal Novae observed in recent years were Nova Aquilae in 1918, Nova Cygni in 1920, Nova Pictoris in 1925, Nova Herculis in 1934, Nova Lacertae in 1936. There have been numerous faint Novae observed in the spiral nebulae: these latter have one advantage over Galactic Novae; the distances of the spirals are now determinable by the Cepheid variables in them, and in other ways, whereas the distance of galactic Novae is difficult to find; they are generally too remote to have sensible parallaxes. Dr Hubble assigned $-5{\cdot}7$ as the mean absolute magnitude of the Novae in the Andromeda nebula. The assumption of this figure for the Galactic Novae would probably give a fair approximation to their distance; other

methods had given $-7{\cdot}4$ as a mean for the Galactic Novae; considering the difficulty of the problem, the two values are not too widely different.

The great interferometer mounted on the 100-in. telescope at Mt Wilson has made it possible to measure the diameters of the bright red giant stars. Betelgeuse in Orion was found to have a diameter 300 times that of the Sun, Antares 450 times, Alpha Herculis 400 times. These values agreed well with estimates previously made by Prof. Eddington from their total light and their surface brilliance, the latter being deduced from their colour. The interferometer made it possible to determine all the elements of the orbit of Capella, the duplicity of which was known by its composite spectrum, but its components are too close to separate in the telescope.

The gaseous nebulae. Huggins was the first to show that the objects classed as nebulae belong to two entirely different categories; the gaseous nebulae, the spectra of which show only a few bright lines, and the extra-galactic nebulae, which are now known to be independent stellar systems at immense distances. The sharpness of the lines in the former class of nebulae indicates that the gases are of extremely low density. The lines of hydrogen and helium were identified, but there were other lines, forming a pair in the green, and another in the ultra-violet, of which the origin was unknown, and a hypothetical element nebulium was suggested as their source. I. S. Bowen deduced that these lines are due to ionized and doubly ionized oxygen, in a metastable state, as a pressure lower than any that can be produced in the laboratory. It is thought that these nebulae are not self-luminous, but are lit up by stars adjacent to them. These stars are generally the intensely luminous stars of types O or B. Thus each of the brighter stars in the Pleiades is surrounded by nebulosity; a larger but fainter nebula envelopes the whole cluster. The stars R. Monocerotis and R. Coronae Australis are irregular variables, and the nebulosity surrounding them seems to vary in sympathy. The distance of the Great Nebula in Orion is about 600 light-years; the spectroscope appears to indicate its rotation in a period approaching a million years, implying a mass ten thousand times that of the Sun.

Expanding nebulosity is generally seen after the outburst of a Nova, obviously composed of the gases driven off from the explosion. But in some cases the expansion is too rapid for this, and implies that the gases were already there, but only rendered visible when the light of the Nova reached them. Expansion has lately been detected in the Crab Nebula in Taurus, suggesting that it may be the relic of a former Nova; a Nova is stated to have appeared in this region in July 1054, which would agree with the observed expansion.

Extra-galactic nebulae. It is in the regions beyond our own star-system that the greatest advance of knowledge has been made in recent years. The nearest objects in these regions are the two Magellanic Clouds, which may be considered distant satellites of the Galaxy, analogous to the two companions of the Great Andromeda Nebula. It was from the Cepheids in the smaller Cloud that Miss Leavitt discovered in 1908 the law connecting period of variation with absolute magnitude; from this law Shapley found the distance of the Large Cloud to be 85,000 light-

years, of the Small Cloud 95,000; the distance between them is about 35,000; their diameters are 12,000 and 6000 light-years. Till their distances were determined it was doubtful whether the Clouds belonged to the Galaxy or not; they are found to contain the same classes of objects as the Galaxy, though without dwarf stars since stars intrinsically faint cannot be seen at such a distance. The Large Cloud contains an immense gaseous nebula, and a variable star which has a mean absolute magnitude of $-8{\cdot}3$, and at maximum is 300,000 times brighter than the Sun. The total luminosities of the two Clouds are 575 million and 230 million times that of the Sun, respectively.

The next object that we meet in our outward journey is N.G.C. 6822, discovered by Barnard in Capricornus; it is a faint irregular nebula, somewhat resembling the Magellanic Clouds, but much smaller and fainter; it is only 20° south of the Galaxy (in Aquila) so its light is assumed to be dimmed by the obscuring clouds which fringe the Galaxy; this has been allowed for in estimating its distance from 12 Cepheids found in it; it is taken to be 530,000 light-years; its outline is elongated, the length and breadth being 3200 and 1600 light-years; it has a central core, extending over one-third of its length and breadth. This nebula ranks as an extreme dwarf, one of the faintest known; the next object on the list is very different; the Great Nebula in Andromeda, Messier 31: this is easily visible to the naked eye; Webb in his *Celestial Objects for Common Telescopes* states that a Persian astronomer Sufi, in the tenth century, mentioned it as a familiar object; the idea that this, and similar objects, are external universes was held by Sir William Herschel, and was still prevalent in the first half of the nineteenth century; but the spectroscopic proof by Huggins that many nebulae are gaseous raised doubts as to the status of the others. Sir David Gill and others revived Herschel's view; doubt was set at rest by a study by Hubble of many Cepheid variables in the outer portions of the great spiral nebulae in Andromeda and Triangulum. He established that the Cepheids belonged to these nebulae by finding none in the regions at some distance outside the nebulae. The Leavitt law connecting period of light-change with absolute magnitude indicated distances of the order of 900,000 light-years for these nebulae; subsequently allowance for absorption of light by dust-clouds round the Galaxy reduced the distance to 680,000 light-years; Isaac Roberts's photographs had already shown that these nebulae are spirals, and probably they give an idea of what the Galaxy would look like at that distance; two dark streaks (the spaces between the spiral arms) had been already detected in the Andromeda nebula by Bond. The Andromeda nebula has two companions, which are analogous to the Magellanic Clouds. Also Hubble has found some star-clusters adjacent to it; these are, however, much smaller than those in the Galaxy. More than a hundred Novae have been seen in it, most of them comparable with Galactic Novae in absolute magnitude; but that of 1885, near the centre of the nebula, reached the seventh magnitude, implying a lustre 100 million times that of the Sun. The term Supernova is used for such objects, which have not been recorded in the Galaxy, with the possible exception of Tycho's star of 1572 in Cassiopeia. The Andromeda Supernova caused underestimates of the distance of the Great Nebula to be accepted till

Hubble derived the true value. Spectroscopic photographs of this nebula showed that it is rotating, which fact helped in the recognition of Galactic Rotation; indeed some such motion appears necessary to avoid collapse, as in the planetary system. No separate star-images are discernible in the two companion nebulae of the Great one; they probably consist mainly of dwarf stars; their spectra confirm this, showing dwarf characteristics.

Only 15° distant from the Andromeda nebula there is another very conspicuous one, Messier 33 in Triangulum; its distance, measured by Cepheid variables, is very slightly larger than that of M. 31; their distance apart is only a quarter of that from our system. We see M. 33 less foreshortened than M. 31; it has a bright centre with about ten curved streams radiating out from it. Its total mass is estimated as 1000 million suns. Novae have been found in it, also variables of different types. The general tendency for the distant nebulae is to show recession, their spectral lines being shifted towards the red; but these two great nebulae show approach; a large portion of it can, however, be explained by the Sun's revolution round the Galactic centre, which happens to be more or less towards them.

The spiral form had already been visually detected by Lord Rosse in Messier 51 in Canes Venatici: that form is now known to be very common among the distant nebulae; it would result from a combination of rotational and radial motions. Dr Hubble in *The Realm of the Nebulae* gives a diagram of the suggested stages of development, beginning with a spherical nebula, which through rotation becomes elongated; in later stages the development has two alternative courses; (1) the normal spirals and (2) the barred spirals; the former throw out streams from the two ends of the long axis, which curve round into the spiral shape; the latter begin with a straight bar through the nucleus; from the ends of the bar streams proceed, in some cases outward in the same direction as the bar, in others at right angles to the bar. The spiral streams generally contain brighter portions, probably star clusters, at intervals; also many of them have bands of obscuring material round them, resembling the obscuring clouds in the Galaxy, such as the "Coal Sack".

The method of finding distances of nebulae by observing the periods and the apparent magnitudes of Cepheid variables is available only for the nearer nebulae. For those farther away, but in which star-images can be seen, the following method is used: in the nebulae with visible Cepheids the absolute magnitude of the fifth brightest star in each nebula is found; the very brightest stars are not used, as they may be abnormal stars; then the distances of the nebulae without Cepheids are found from the apparent magnitudes of their fifth brightest stars. For nebulae too distant for any stars to be seen the apparent magnitude of the nebula itself gives a rough measure, but not very exact, since there is a considerable range in the absolute magnitudes of nebulae. Fortunately at this point another criterion of distance becomes available. It has already been mentioned that the bright nebulae in Andromeda and Triangulum show motion of approach; this is, however, chiefly due to the Sun's revolution about the Galactic centre; when more distant nebulae are examined spectroscopically (a spectroscope of short focus and great light-grasp

has been designed for the purpose at Mt Wilson), the shift of the spectral lines is invariably towards the red, indicating recession; and there is close correlation between the red-shift and the magnitude, so that the red-shift is an excellent measure of distance; when the distance is great, any peculiar motion of the nebula itself can only be a small fraction of that indicated by the large red shift.

The red shift in the spectra of distant objects had already been predicted by Prof. de Sitter in 1916, long before the evidence from the distant nebulae was available; the existence of the shift is now universally admitted, the only uncertainty being whether it arises from actual expansion of the universe, or from some unknown property of space. Dr Edwin Hubble, who took part in the photographic research on the nebulae undertaken with the 100-in. reflector at Mt Wilson, and has subsequently discussed the results in two books, *The Realm of the Nebulae* and *The Observational Approach to Cosmology*, inclines to the latter alternative; one of his reasons is the comparatively short time-scale that the former alternative involves. It would mean that all the nebulae, which now fill such an immense volume, were in fairly close proximity to each other about 2000 million years ago; this is a shorter period than the geologists assign to the oldest rocks on the Earth's surface. It also implies that space is limited, re-entering on itself, and that a large fraction of it is already included in the survey made by the 100-in. reflector; the 200-in., now in process of construction at Mt Palomar, California, will penetrate twice as far into space, and should help materially to a decision. The estimated distance of the most distant nebulae reached at present is about 400-million light-years; the immense number of them is shown by a photograph reproduced by Dr Hubble of a field near the north galactic pole; this is so thickly crowded with images as to form an almost continuous background; it is stated that about half of them belong to stars in our system, the other half to external nebulae.

A century ago O. M. Mitchell wrote[5]: "For every star visible to the naked eye the telescope reveals a universe, far sunk in the depths of space, and scattered with vast profusion over the entire surface of the heavens." This remains true to-day, only we may change "visible to the naked eye" into "visible with a powerful telescope"; also we must insert the words "apart from the Milky Way" at the end.

Dr Hubble gives another reason for his hesitation in accepting the red shift in the spectra of distant nebulae as due to recession. He points out that the red shift in any case implies a reduction in the energy received, since the region of maximum energy in the spectrum is cut off. But if there is recession there will be a further reduction in the energy received, since the recession causes fewer quanta to reach the observer per second. This loss is analogous to the lowering of the note of the whistle when a whistling locomotive passes an observer. Dr Hubble claims, from the statistics now to hand, that the observed counts of nebulae of different magnitudes and different red shifts accord better with the single reduction of energy than with the double reduction. But he does not claim that this verdict is final, and all astronomers look forward with keen interest to the completion of the 200-in. telescope, while expressing their admiration and gratitude for the splendid results that have already been obtained with the 100-in.

NOTE ADDED IN PROOF

The author having been incapacitated by an accident, the proof was corrected by the Rev. Dr M. Davidson, who added the following note.

Two new satellites of Jupiter were discovered by Dr S. B. Nicholson on photographs made with the 100-inch telescope, Mount Wilson, in July. They were both of magnitude about 19 and must be very small bodies, probably about 10 miles in diameter. The nearer to Jupiter of the two is revolving in an orbit with eccentricity 0·140 at an inclination of 28° in a period of 254 days. Its mean distance from Jupiter is over 7 million miles, so it belongs to the same group as satellites VI and VII. The outer satellite has a retrograde motion, its inclination being 163°, its eccentricity 0·207, and its mean distance is 14 million miles. It revolves around its primary in 693 days. It is remarkable that this satellite should also belong to another group comprising satellites VIII and IX, being only a little nearer to Jupiter than these.

REFERENCES

(1) JEANS, Sir J. In *Encyc. Brit.* (1929).
(2) RUSSELL, H. N. *The Solar System and its Origin*, p. 49 (1935).
(3) JEFFREYS, H. *The Earth*, p. 230.
(4) JEFFREYS, H. *The Observatory*, p. 36 (Feb. 1938).
(5) MITCHELL, O. M. *The Orbs of Heaven.*

METEOROLOGY

By D. BRUNT, M.A., B.Sc.

§ 1. CONDENSATION IN THE ATMOSPHERE

(*a*) *Condensation on nuclei in unsaturated air.* A general discussion of problems connected with condensation in the atmosphere has been given by Bennett[1]. As condensation in the atmosphere always occurs on hygroscopic nuclei, the condition in which it begins to form is not the existence of relative humidity of 100 per cent, but a vapour pressure exceeding the equilibrium vapour pressure over the nucleus. J. S. Owens[2] observed occasions when nuclei began to gather moisture at a relative humidity of 74 per cent. Pick[3] showed that fogs (not smoke fogs) can, and most frequently do, occur with unsaturated air, not only at the sea coast, but also inland. Sohoni and Paranjpe[4] found that very large numbers of thick fogs in India occur with relative humidities of 90 per cent or less, even at country stations where the atmosphere is not dusty.

(*b*) *Sizes of water drops in fog and cloud.* The effectiveness of the coagulation of small drops to form larger drops, in fog, cloud and rain, appears to be negatived by hydrodynamical theory, but Köhler[5] observed that the masses of droplets in fog and cloud are integral multiples of a certain unit, and that the concentrations of salt in solution in the drops are integral multiples of a certain unit concentration. These observations, and those of Defant[6] and Niederdorfer[7] that the masses of raindrops are integral multiples of a standard drop, appear to demand the formation of large drops by the union of drops of a standard size. Kneusel[8] quotes for the mean size of cloud droplets values varying from 2 to 8 times 10^{-4} cm.

(*c*) *Formation of raindrops.* Bergeron[9] suggests that the direct condensation of water vapour on to cloud droplets is never sufficient to produce drops large enough to fall as rain. He rejects the possibility of coagulation as a factor in the growth of drops, and puts forward the theory that rain is formed by the action of ice crystals falling through a cloud of small water drops. The vapour pressure being lower over an ice crystal than over a water drop, the liquid water evaporates and condenses on the ice, and this, in falling to the ground, melts into rain. Findeisen[10] finds, from an examination of observations made on aircraft over a period of some years, that rain of even moderate amount, and particularly rain in which the drops are large, originates in nearly all cases as ice condensation.

(*d*) *Supercooling of water in the atmosphere.* Findeisen[10] states that at cloud temperatures just below freezing point water drops predominate, while at cloud temperatures below $-10°$ C. ice particles predominate. Clouds of liquid drops have, however, been observed at temperatures as low as $-40°$ C. The increase of flying of civil aircraft in winter conditions has directed attention to the freezing of supercooled water drops on impact with the leading edge of the wing and other parts of aircraft. The meteorological conditions which favour the formation of ice on aircraft

have been summarized by Bigg[11], Simpson[12], Noth and Polte[13], and Byers[14]. Bleeker[15] maintains that an important role in icing is played by the evaporation from the water-ice mixture.

(*e*) *Hail formation.* The theory of hailstone formation has been discussed by Schumann[16], who lays emphasis on the evaporation from the surface of the hailstone during the period of growth as a partial compensation for the latent heat liberated by the freezing of water on its surface.

Bilham and Relf[17] have shown from the relationship between terminal velocity and diameter of a sphere that there is an upper limit of about 1·5 lb. to the mass attainable by a spherical hailstone. This corresponds to a diameter of about 5 in. Beyond this limit of size the terminal velocity increases so rapidly that a sufficiently strong upward current to maintain the hailstone in the air can rarely occur.

§ 2. THE FORMATION AND DISSIPATION OF NON-FRONTAL CLOUDS

The conditions in which clouds with a cellular or roll pattern occur in the atmosphere were stated by Chandra[18], and slightly amplified by Brunt[19], in continuation of the work done by G. T. Walker and his pupils. Pure cellular motion is produced when the degree of instability surpasses the limit set by the Rayleigh-Jeffreys theory[20]. With very small shear of wind through the layer, the cells are slightly distorted; with greater shear the cells are replaced by long rolls across the direction of shear; and with large shear the clouds are long rolls along the direction of shear. Straight long rolls of cloud are invariably associated with wind increasing with height without change of direction.

When sheets of cloud clear at or soon after sunset, the initial stage is frequently the breaking of the continuous sheet into small cloudlets or rolls, the form being determined by the variation of wind with height[21]. It was suggested by G. T. Walker[22] that all rolls of cloud are instability patterns, but this is by no means certain (see references to J. Bjerknes and Godske in § 8 (*a*) below). Suckstorff[23] found evidence that the clouds from which instability showers fall are of the cellular type, with upward motion in the middle.

Genuine stratus clouds, or raised fogs, are produced by surface cooling combined with turbulence, which raises the fog off the ground; but strato-cumulus cloud, and much of the cloud designated as stratus, can be formed by the joint effect of turbulence and variation of surface temperature. The use of the tephigram or any other thermodynamic diagram makes it easy to estimate the height at which air lifted from the ground will attain saturation, and so form cloud; and if the variation of surface temperature can be foretold with reasonable accuracy, the variation of the height of the cloud can be foretold. Poulter[24], who has given an account of the use of the tephigram for such forecasts of cloud height, points out that while the clouds are formed by condensation in air originating at the ground, the persistence of the cloud after formation will depend on the relative humidity of the environment in which they are formed. If the environment is dry, the clouds

will evaporate quickly. Poulter discusses a number of occasions in detail, including cases of dissipation of cloud as an effect of turbulence combined with surface heating. The latter problem has also been discussed by Krick[25]

§ 3. VISIBILITY

The theory of atmospheric visibility has been outlined by Bennett[26] and Middleton[27]. Wright[28] found that at Kew Observatory during the afternoon the visibility, if above 2 km., is practically independent of the number of nuclei and of the relative humidity, but is determined almost entirely by the number of solid particles in suspension in the atmosphere. Wright found that the distance of vision in these circumstances varied inversely as the number of solid particles raised to a power slightly greater than unity. Low visibilities, and particularly the occurrence of mist and fog, are associated with high relative humidity, and with high concentrations of both solid particles and nuclei of condensation, the density of the fog increasing with each of these three variables.

Bennett[29] has given a table of the distance of vision of lights of various candle powers on the average at night, in meteorological conditions corresponding to different daytime visibilities.

§ 4. APPLICATION OF THERMODYNAMIC DIAGRAMS IN METEOROLOGY

(*a*) *The representation of upper air observations.* When observations of temperature are made at different levels in the atmosphere, the vertical co-ordinate actually observed is the pressure. Thermodynamic diagrams used for the representation of the observations will therefore usually have as their co-ordinates the temperature (T) and the pressure (p), or some functions of these two variables. Or one of the co-ordinates may be a measure of humidity. The diagrams most commonly used are

(i) The Neuhoff[30] diagram, with co-ordinates T and $\log p$.

(ii) The Tephigram[31], with co-ordinates T and the entropy, usually designated by the symbol ϕ. Since entropy is proportional to the logarithm of potential temperature, θ, the co-ordinates may be regarded as T and $\log \theta$.

(iii) The Rossby diagram, with co-ordinates entropy of dry air and the humidity-mixing-ratio[32].

(iv) The Refsdal aerogram, with co-ordinates $\log T$ and $T \log p$[33].

The tephigram is in common use in the British Isles and many other parts of the world. In this diagram the dry adiabatics are horizontal straight lines, and the isothermal lines are vertical straight lines.

(*b*) *The wet-bulb potential temperature.* Normand[34] showed that as a mass of damp air ascends or descends through the atmosphere, the point representing its dry-bulb temperature follows a dry adiabatic line, and the point representing its wet-bulb temperature follows a saturated adiabatic. As soon as the air becomes saturated during ascent the dry- and wet-bulb temperatures coincide, and follow

the saturated adiabatic on which the original wet-bulb temperature lay. The wet-bulb temperature corresponding to the intersection of this saturated adiabatic with the isobar of 1000 mb. is called the wet-bulb potential temperature, and is usually designated by the symbol θ'. It is readily seen that θ' is an invariant for any mass of air during any adiabatic changes, and it has the further merit, as Normand showed, of being invariant even when there is condensation or evaporation taking place within the air mass under consideration. Normally the wet-bulb potential temperature will vary with height in the atmosphere, but during any changes of position in which no heat is given to or taken from the air by radiation or turbulence the value of θ' for any selected mass of air will remain invariant, and so may be used as a means of identifying that mass at any time.

The use of the wet-bulb potential temperature in practical meteorology was first developed by Hewson[33]. The method he used is best shown by means of an

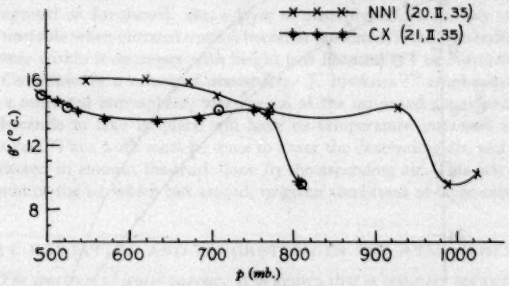


Figure 1.

example. In figure 1, curve *NNI* shows the variation of θ' with pressure as observed at different heights above San Diego on 20 February 1935. An examination of the upper winds over the western half of the United States showed that the air in which these observations were made was approximately over Cheyenne on the 21st. The upper air observations at Cheyenne on that day yielded the distribution of θ' with pressure shown in curve *CX* in figure 1. The similarity of the *CX* and *NNI* curves, with a shift of about 180 mb. in pressure, is very striking, and may be held to corroborate the identity of the air masses in which the observations were made on the two consecutive days, except for the loss of a relatively thin layer of the initial surface air at San Diego, probably diverted by the marked irregularities of the ground.

The use of this method depends on the θ', p curve having some characteristic form. Figure 2, also taken from one of Hewson's papers, represents in the same way a pair of ascents made, so far as it is possible to judge from the upper winds, in the same air mass. Comparing the pressures at which equal values of θ' occur

over the two stations, we find that while the air initially at a pressure of 500 mb. had subsided to approximately 700 mb., the air initially at 700 mb. had reached a pressure of about 820 mb. Thus the subsidence was accompanied by marked divergence, the amount of this divergence being found to agree very closely with the amount computed from the wind distribution, which in this exceptional case was sufficiently clearly defined to make an estimate of divergence possible. Hewson[35. III] used similar reasoning to show that the air ascending over a warm front diverges, as suggested by J. Bjerknes[36].

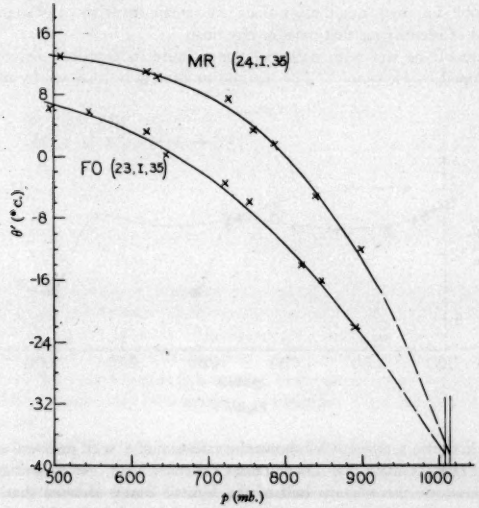


Figure 2.

(*c*) *The general conditions of stability in the atmosphere.* Normand[37] has summarized the conditions for stability in the atmosphere as follows:

A dry atmosphere will be stable, for small or large displacements of individual particles of air, if the potential temperature increases with height, i.e. if the lapse-rate is less than the dry adiabatic.

A saturated atmosphere will be stable, for small or large displacements of individual particles of air, if the lapse-rate is less than the saturated adiabatic.

In an unsaturated atmosphere the conditions for stability or instability, for small displacements, are the same as those for dry air. But when portions of an initially stable unsaturated atmosphere are given *large* displacements, sufficient

to lead to the displaced air becoming saturated, instability may or may not arise, according to the initial distribution of temperature and humidity with height. Normand[37] has shown that if the saturated adiabatic through the point representing the initial wet-bulb temperature of the displaced mass does not anywhere cross the line representing the distribution with height of the dry-bulb temperature, no upward displacement of that mass can bring it to a level in which it will be warmer than its new environment, and so be unstable. If however these two lines intersect, their intersection determines the height beyond which the displaced particle will be unstable, and so will rise freely. In the latter case, the air is said to have *latent instability*.

A *layer* of dry air is stable or unstable according as the lapse-rate within it is less than or greater than the dry adiabatic, and a layer of saturated air is stable or unstable according as the lapse-rate within it is less than or greater than the saturated adiabatic. The state of stability or instability is unaffected by any vertical motion upward or downward. But a layer of unsaturated air, initially stable, will become unstable when elevated until it becomes saturated, if the wet-bulb potential temperature within it decreases with height (see Hewson[35, I] or Normand[37]).

(*d*) *Convection in a saturated atmosphere.* J. Bjerknes[38] emphasizes the fact that, in a saturated atmosphere, when some of the saturated air ascends, the air which descends to take its place will have its temperature increased at the dry adiabatic rate. Thus work must be done to lower the descending air, and this work cannot exceed in amount the work done by the ascending air. This sets a limit to the amount of the air which can ascend, in given conditions of lapse-rate.

§ 5. RADIATION AND ABSORPTION IN THE ATMOSPHERE

(*a*) *The spectrum of water vapour.* It is known that at ordinary temperatures the normal constituents of dry air (except carbon dioxide) do not absorb or radiate an appreciable amount of energy[39]. The only constituent of the atmosphere whose absorption is great is water vapour, which absorbs and radiates within a wide range of wave-lengths. Carbon dioxide has an intense but narrow band of absorption centred at $14{\cdot}7\,\mu$.

Most of the computations of radiation and absorption in the atmosphere have been based on Hettner's[40] observations of absorption in columns of steam. More recent observations by Weber and Randall[41], in water vapour at lower temperatures, yield coefficients of absorption much lower than those of Hettner. A comparison of the two series of observations has been given by Wexler[42]. The general question of the effect on the form of the lines in the water-vapour spectrum of variations in the temperature and the partial and total pressure has been discussed by Elsasser[43], on the basis of Dennison's[44] theory of the form of a spectral line. It is however still doubtful what is the true nature of the spectrum of water vapour, and how it varies with temperature and total and partial pressure. Albrecht[45] has suggested that in the conditions which hold in the stratosphere the absorption spectrum of water vapour is reduced to a series of narrow lines.

On the basis of Hettner's data, Simpson[46] assumed that a layer containing 0·3 mm. of precipitable water would (*a*) absorb completely all incident radiation between 5·5 and 7 μ., and above 14 μ., (*b*) be completely transparent to all radiation between 8·5 and 11 μ., and below about 4 μ., and (*c*) absorb incompletely in the ranges 7 to 8·5 and 11 to 14 μ.

On the basis of the above assumptions Simpson obtained reasonably good agreement between the observed downward radiation from the atmosphere and the observed temperature and vapour pressure.

Brunt[47] extended Simpson's argument to show that the transfer of heat by absorption and radiation is closely analogous to the conduction of heat in a solid, the effect in the atmosphere being intermediate between the effects of molecular conduction and turbulence. The numerical results of this work, as well as such computations as those of Möller and Mügge[48] and of Baur and Philipps[49] of the flow of energy by radiation in the atmosphere, are subject to some doubt, on account of the uncertainty as to the validity of the coefficients of absorption of water vapour deduced from Hettner's work in the conditions of temperature and pressure which hold in the atmosphere.

(*b*) *Some empirical results.* Brunt[50] obtained an empirical formula for the downward radiation from the atmosphere in the form

$$E/\sigma T^4 = a + b\sqrt{e},$$

where E is the downward radiation, σT^4 is the total black-body radiation at the surface temperature T, a and b are constants, and e is the vapour pressure. This formula, with two adjustable constants, appears to give better agreement with observations than the formula of Ångström,

$$E/\sigma T^4 = A - B\,10^{-\gamma e},$$

which has three adjustable constants A, B and γ.

Brunt[50] gave for the variation of temperature after sunset, with a clear sky, the following formula:

$$T = T_1 - \frac{2}{\sqrt{\pi}} \frac{R_N}{\rho c \sqrt{\kappa}} \sqrt{t},$$

where R_N is the net loss of heat by radiation, ρ, c and κ are the density, specific heat and diffusivity of the ground, and t the time since sunset.

Cloud sheets radiate effectively as black bodies, the upper surface losing heat on balance, as does the surface of the earth on a cloudless night. The lower surface of a cloud sheet gains heat on balance, and the tendency to produce instability within the cloud layer is greatest when the sum of the gain at the lower surface and the loss at the upper surface is greatest. Durst[51] estimated this to occur at about the level at which Alto-Cumulus cloud occurs, the form of this cloud frequently suggesting instability flow (see § 2 above).

§ 6. TURBULENCE IN THE ATMOSPHERE

(*a*) *Diffusion by continuous movement.* The theory of diffusion by continuous movement was first outlined by G. I. Taylor[52], in terms of the coefficient of correlation, R_ξ, between the velocity of a selected particle of air at time t and

its velocity at a time ξ later. This theory has been developed further in two papers by O. G. Sutton[53], on the assumption that the form of R_ξ is given by

$$R_\xi = \left(\frac{\nu}{\nu + \overline{w'^2}\xi}\right)^n,$$

where ν is the kinematic coefficient of viscosity, n is a parameter whose value is defined by the meteorological conditions at the time, and $\overline{w'}$ is the vertical component of the eddying velocity. Assuming a Maxwellian distribution of the eddy velocities,

$$\overline{w'^2} = \tfrac{1}{2}\pi\,(\overline{w'})^2 = \tfrac{1}{2}\pi l^2 \left(\frac{\partial \bar{u}}{\partial z}\right)^2,$$

where l is the mixing length, $\bar{u}$ the mean horizontal velocity in the direction of mean motion, and z is the vertical co-ordinate. Adopting v. Kármán's form

$$l = k \frac{\partial \bar{u}}{\partial z} \Big/ \frac{\partial^2 \bar{u}}{\partial z^2},$$

Sutton finds for the eddy diffusivity K the form

$$K = (0{\cdot}251)^{1-n} \nu^n \left\{\left(\frac{\partial \bar{u}}{\partial z}\right)^3 \left(\frac{\partial^2 \bar{u}}{\partial z^2}\right)^{-2}\right\}^{1-n},$$

thus relating the eddy diffusivity to the form of the velocity-height curve.

Also assuming the eddy stress not to vary with height, Sutton finds for the variation of wind with height the equation

$$\bar{u}/\bar{u}_1 = (z/z_1)^{n/(2-n)},$$

a relationship which is found to fit experimental data with a high degree of accuracy. The parameter n is thus directly derivable from the velocity-height curve. For $n = \frac{1}{4}$, $\bar{u}/\bar{u}_1 = (z/z_1)^{\frac{1}{7}}$, the law which represents the velocity-height curve in wind tunnels for a wide range of conditions.

(*b*) *The logarithmic law of variation of wind with height.* Rossby and Montgomery[54], assuming that the mixing length l may be represented by

$$l = k\,(z + z_0), \qquad \text{.....(1)}$$

where k is a constant (0·38 according to Prandtl and v. Kármán), found that if the eddy shearing stress, $\widehat{xz}$, does not vary with height, the mean wind at height z is given by

$$\bar{u} = \frac{1}{k}\sqrt{\frac{\widehat{xz}}{\rho}} \log_e \frac{z + z_0}{z_0}, \qquad \text{.....(2)}$$

or

$$\widehat{xz} = \rho k^2 \bar{u}^2 \Big/ \left\{\log_e \frac{z + z_0}{z_0}\right\}^2. \qquad \text{.....(3)}$$

$$K = k^2 (z + z_0) \frac{\bar{u}_a}{\log_e \dfrac{z_a + z_0}{z_0}}, \qquad \text{.....(4)}$$

where K is the eddy diffusivity, and $\bar{u}_a$ the wind at the height z_a of the anemometer.

The logarithmic law of equation (2) has found many applications in wind-tunnel experiments as well as in Meteorology. Values of K deduced from equation (4) also show close agreement with values deduced by Mildner[55] from wind observations.

The above equations are only valid within the range of height in which it is permissible to regard the shearing stress as constant with height. The constant z_0 is represented as equal to $s\epsilon$, where s is a constant and ϵ is the linear dimension of the irregularities of the surface of the ground. Prandtl suggested a value $s = 1/30$, but this was based on experiments in which there was a close relation between the size and packing of the surface inequalities, and this value is not of necessity applicable in the atmosphere. Rossby and Montgomery suggest the following values of z_0 over different types of surface:

Over short cropped grass	...	$z_0 = 0{\cdot}5$ cm.
Over open grassland	...	$z_0 = 3{\cdot}2$ cm.
Over sea with small swell	...	$z_0 = 3{\cdot}9$ cm.

Equation (3) above is of interest, in that $\widehat{xz}$ is equal to the skin friction at the ground. G. I. Taylor[56] suggested that the skin friction could be represented by

$$\rho\kappa\overline{u}^2, \qquad \text{.....(5)}$$

where κ is a constant, whose value is 0·004 in wind-tunnel experiments. In the atmosphere Taylor found κ to vary from 0·0022 to 0·0032. The two forms (3) and (5) above are identical if

$$\kappa = \frac{k^2}{\left\{\log_e \frac{z_a + z_0}{z_0}\right\}^2}. \qquad \text{.....(6)}$$

With $z_0 = 3{\cdot}2$ cm., $z_a = 30$ m., this equation gives $\kappa = 30 \times 10^{-4}$, in close agreement with the upper limit derived by Taylor from wind observations over Salisbury Plain.

The values of κ deduced by Sutcliffe[57], using winds at lower heights than Taylor, were rather higher than the values deduced by Taylor, and Durst[58] showed that Sutcliffe's data could be rearranged to show that κ is much greater in conditions of temperature inversion than in lapses. These results cannot strictly be discussed on the basis of the above theory, which assumes an adiabatic atmosphere, but it is of interest that Sutton suggested[59] that the influence of surface inequalities is small in large lapse-rates and large in inversions.

In a stable atmosphere, Rossby and Montgomery found for the variation of wind with height the differential equation

$$\frac{\partial \overline{u}}{\partial z} = \frac{1}{k(z+z_0)}\sqrt{\frac{\widehat{xz}}{\rho}}\sqrt{\left[\tfrac{1}{2} + \tfrac{1}{2}\sqrt{\left\{1 + \frac{4\kappa^2(z+z_0)^2}{\widehat{xz}/k^2\rho}\right\}}\right]},$$

which is only applicable to an atmosphere in which the lapse-rate is constant.

Sverdrup[60] extended this theory, assuming, instead of a constant lapse-rate, that near the surface the variations of velocity and temperature follow the same law, and derived a number of relationships which gave very good agreement with obser-

vations made over a snow surface. This method has the merit that it separates out the effects of stability and of surface roughness.

(*c*) *Eddy diffusivity for different atmospheric properties.* It has been suggested that the numerical value of the eddy diffusivity might differ for heat, momentum, content of water vapour, etc. H. Berg[61] showed that certain observations appeared to require lower values for humidity than for temperature. It is however difficult to ensure that the effects of radiation and absorption on the distribution of temperature are adequately allowed for.

(*d*) *Evaporation.* Sutton[53] has treated the problem of evaporation as a two-dimensional problem, and found that the evaporation from a strip of length x_0 downwind, and unit width across wind, assuming no edge losses due to lateral diffusion, is given by

$$E = B\bar{u}^{(2-n)/(2+n)}\, a^{2/(2+n)}\, x_0^{2/(2+n)}.$$

Here a is a quantity whose value depends on n and on the constants of the atmosphere, but is independent of the mean wind $\bar{u}$. The parameter n is determinable from the velocity-height curve.

For a circular area of radius r,

$$E = B'\bar{u}^{(2-n)/(2+n)}\, a^{2/(2+n)}\, r^{(4+n)/(2+n)}.$$

For turbulent flow in a wind tunnel it is known that the velocity distribution over a smooth plane surface is given very accurately, for a wide range of Reynolds numbers, by

$$\bar{u}/\bar{u}_1 = (z/z_1)^{\frac{1}{4}},$$

which requires $n = 1/4$ on Sutton's theory. It follows that the rate of evaporation from rectangular or circular areas should vary as $\bar{u}^{0.78}$, whilst the variation with length of the surface downwind is given by $x_0^{0.89}$.

A comparison of this theory, as regards the variation with velocity, was given by Lettau[62], who quotes the results of Dorffel, yielding an index 0·71, and those of Mrose, made with similar apparatus, yielding an index 0·75. Himus found the index 0·77, and Sutton[53] analysed Hine's measurements of evaporation with liquids other than water, showing that there was good agreement with the index 0·78. Powell and Griffiths[63], using a somewhat unexpectedly large value for the evaporation in a calm, gave a value 0·85 for the index, but if the evaporation in the absence of wind be treated as negligible, their observations give a good fit with the index 0·78.

In any discussion of measurements of evaporation made with evaporimeters it must be borne in mind that the evaporation will be influenced chiefly by the eddies which are not greater in size than the dishes used, so that the turbulence set up by the dishes themselves will tend to be the controlling factor. Such measurements will therefore not reflect the major diurnal changes of wind structure, but will always tend to approximate to the wind-tunnel type of evaporation. Millar[64] found that circular dishes of diameter 50·8 cm. and 11·2 cm. yielded indices 0·77 and 0·725, respectively.

As regards the variation of evaporation with the downwind dimensions of the surface, Lettau and Dorffel[65] found the law $x_0^{0.875}$. Powell and Griffiths found the index to vary with the cross-wind dimension, showing that in their experiments there was appreciable lateral diffusion. These two workers also found a marked effect of the edge of the dish, when this was raised above the level of the evaporating liquid.

It is of interest to note that Elias[66] found for the rate of loss of heat from a rectangular plate the law $\bar{u}^{0.8} x_0^{0.89}$, in practically perfect agreement with Sutton's theory of evaporation, thus indicating that the transfer of heat and matter follow the same law.

§ 7. MICROCLIMATOLOGY

The earlier work on microclimatology is summarized by Geiger[67]. It is found that temperature varies considerably within short distances, depending on the type of soil, soil covering, elevation, slope of ground, attitude of ground relative to the sun, and the general configuration of the ground around. Some examples of the effects of these factors have been given by Schmidt[68], Hawke[69], Cornford[70], and by Ramdas in India[71].

Some of the results which have been obtained are of great importance in relation to the bearing of site on liability to frost damage. On long slopes the effect of night cooling, when the sky is cloudless and the wind light, is to produce a downward flow of the cold air, known as a katabatic wind. Such cold air follows the slopes of the ground in the same way as a stream of water will. The lowest temperatures are usually found at the bottom of the slopes, or on flat ground at the foot of a slope, when there is no outflow for the cold air coming from the higher ground. The ideal site for, say, an orchard, is therefore one which is so situated that (*a*) there is no large area of ground at a higher level from which cold air can drain down to it, and (*b*) where there is a slow downward slope capable of draining away any air cooled on the site. The downflowing cold air can be dammed by a thick hedge or wall, until the depth of the cold air surpasses the height of the obstacle, after which it overflows the latter.

A detailed investigation of katabatic winds in a valley was described by Heywood[72], and an investigation, not yet completed, has been described by Cornford[70].

§ 8. THE DEPRESSION OF MIDDLE LATITUDES

(*a*) *The wave theory of cyclones.* Since the publication in 1922 of the paper by J. Bjerknes and H. Solberg on the *Life Cycle of Cyclones*, advances in the understanding of the depression of middle latitudes have been slow, rather of the nature of the dotting of i's and crossing of t's than of marked advances. The association of the birth of the depression with the surfaces of separation between warm and cold currents in the lower troposphere led to the view that the cause of the origin of the depression is to be found in the lower troposphere, the prime cause being assumed to be the formation of an unstable wave-like disturbance in the surface of separation (the polar front surface) between the two currents, but the earlier

attempts to produce a theoretical basis for this view were not very successful. A mathematical theory has now been developed by H. Solberg[73]. A simplified account of the theory has been given by J. Bjerknes and C. L. Godske[74]. These authors point out that waves at a surface of discontinuity may be affected by (*a*) compressibility, (*b*) gravitation, (*c*) inertia due to the earth's rotation, and (*d*) shear of wind across the surface. They consider these factors in turn, and find that a combination of factors (*c*) and (*d*), i.e. shear of wind across a surface of discontinuity, on a rotating earth, can give rise to waves whose nature depends on their wave-length. Waves whose lengths are of the order of a few kilometres are instability waves, the amplitude of which will therefore increase until they are observable. Larger waves, of which the wave-length is less than about 1000 km., will be stable waves, whose amplitude will therefore not grow beyond the initial value imposed by the disturbance which produces them, and which will therefore not become observable. Waves which have a wave-length of the order of 1000 km. or more will be unstable waves, and will grow in amplitude and become observable. Beyond a few thousand kilometres there is a further range of stable waves. We are thus led by dynamical theory to the conclusion that at a surface of discontinuity on a rotating earth there can occur a class of waves of the right type, and with the right wave-length, to account for the genesis of the depression, as well as short instability waves of the order of magnitude of the so-called Helmholtz waves, deduced by Helmholtz from the classical theory of hydrodynamics, to explain the clouds which take the form of long parallel billows. Small wave-like disturbances which often occur on microbarograph records can be shown, by comparison with wind records, to be due to disturbances of a few kilometres in wave-length, and these may possibly be the shorter type of disturbance described by Bjerknes and Godske.

The mathematical theory is far from complete. The waves it discusses are infinitely long waves, whereas in the depression of middle latitudes the gradients of pressure are of the same order of magnitude in all directions from the centre. The theory is only applicable to the initial stage of development of the depression, and has not been extended to the later stages of occlusion of the depression.

In a detailed study of the kinematical features of depressions, Goldie[75] reached the conclusion that both in the forward and the rear parts of a depression, and notably in the cold air, the average surface motion is approximately that of a simple vortex (vr = constant). This suggests that convergence of air towards the centre is the predominating factor in the growth of cyclonic circulation. There is a central area of polar air within which the velocity is proportional to the distance from the centre, but the warm air which approaches the centre shows no such relationship of velocity to distance from the centre.

(*b*) *The high-level disturbances in depressions.* J. Bjerknes[76] and J. Bjerknes and Palmén[77] have described in detail the structure of the tropopause above a number of selected depressions. Bjerknes infers that the sinusoidal deformations of the tropopause shown by the observations are to be explained by horizontal rather than by vertical motions, a crest on the wave-like structure of the tropopause being a region to which air has moved from lower latitudes.

In a later paper J. Bjerknes[78] discussed the flow associated with a train of wedges of high pressure separated by troughs of low pressure. He found that the effect of gradient wind flow is to produce convergence east of, and divergence west of, a ridge of high pressure, giving a rise of pressure east of, and a fall of pressure west of, the central line of the ridge, the effect being that of a translation of the ridge eastward. A similar translation is derived for a wedge of low pressure.

The Norwegian school of meteorologists regard the perturbations at the tropopause level as being started by, and travelling with, the polar front wave which produces the depression as observed in the lower troposphere. The polar front wave is regarded as the primary wave, and the tropopause disturbance as a secondary effect. This is in direct conflict with the views of the German school of thought, according to which the stratospheric disturbance is the primary phenomenon, and the tropospheric disturbance a secondary phenomenon. A clear summary of the views of the Frankfurt and Leipzig schools has been given by Schmiedel[79]. Mügge[80] summarizes the situation by saying that the trough of low pressure in the stratosphere is not the cause, but the effect of the processes in the central part of the depression, these processes being of the nature of a pumping out in this region. He states that it is over the nature of this pumping out that differences of opinion arise. He criticizes Godske's theory regarding the effect of the shearing instability at a front as not applicable in practice. Mügge adduces Scherhag's theory[81] that high-level falls of pressure are frequently compensated by the inflow of cold air below, which has the effect of producing a front as a *consequence* of the depression, not as its *cause*. He envisages the free unstable wave in a depression as being essentially in the free atmosphere, the fall of pressure being greatest at high levels, and decreasing downward as a result of the compensation described by Scherhag, many high-level depressions being in fact not appreciable at the ground. Mügge maintains that fronts are more clearly defined *during* than *before* the development of the depression.

Other papers of interest in regard to the differences between the views of the German and Norwegian schools are those of Ficker[82], Mollwo[83], Raethjen[84], and Mügge[85].

The Leipzig school regards much of the weather as being dominated by a combination of periodic terms, and it is found that, at least in the lower troposphere, the amplitudes of some of these terms increase with height. From this it is inferred that the primary cause is to be found at high levels.

This aspect of meteorology is as yet scarcely ready for statement in textbook form. Further progress demands the study of more individual depressions, as regards the distribution of temperature, humidity and wind. It is probable that large differences of structure and mode of development will be found to subsist between different depressions, and that whilst some atmospheric disturbances are primarily of tropospheric origin, others may be primarily of stratospheric origin. But phenomena in the troposphere and stratosphere must be coupled together in some manner.

The development of the counter-clockwise circulation of winds in the lower

troposphere in a depression demands the convergence of the air inward towards the centre, and if the pressure is to fall at the surface there must be divergence at high levels more than sufficient to compensate for the inflow at low levels. Scherhag[86] found that the divergence of winds at high levels is always associated with the development of a depression. The converse proposition was shown by Köppen[87] to hold for anticyclones. Thomas[88], in an examination of the contributions of different levels in the atmosphere to the surface changes of pressure, found that the greatest contribution is from about 10 to 14 km. Scherhag's[89] divergence theory of depressions demands as a necessary pre-condition a horizontal divergence of the isobars, and hence of the stream of flow, at high levels. He finds that the effectiveness of a front in producing changes of pressure is proportional to $(\Delta T)^2 \tan \alpha$, which is approximately proportional to the product of the differences of temperature and wind across the front. Here ΔT is the difference of temperature in the two currents, and $\tan \alpha$ the slope of the surface of separation.

Bjerknes[90] associated the divergence at high levels with the ascent of warm air over the cold air at a warm front, resulting in the development of an anticyclonic circulation at high levels. That this divergence actually takes place was demonstrated by Hewson[91] by the use of wet-bulb potential temperatures for the identification of the air. The divergence of the air at cirrus levels, which is above the levels considered by Hewson, was brought out by Douglas[92] in a diagram of winds at cirrus levels in a depression. This is confirmed in a similar diagram by Palmén[93].

(*c*) *Rainfall in depressions.* Douglas[94] points out that heavy rain is to be observed frequently in warm air close to a practically stationary front, and asks why the warm air converges and rises. In the bodily ascent of such large masses of air as are involved in the production of this rain the lapse-rate of temperature cannot be the deciding factor. Heavy rain is in fact frequently associated with marked stability. A distinction must be drawn between heavy rain in a depression and thunderstorm rain, the latter being largely determined by thermal instability.

§ 9. LONG-RANGE WEATHER FORECASTING

During the last few years a trial has been made in Germany of the issue of forecasts for ten days ahead, the work being under the superintendence of Dr Franz Baur. A considerable degree of success has been attained by these forecasts. The principles on which they are founded have been described by Baur[95], and summarized by Walker[96].

The changes in the surface pressure during a period of 24 hr. are almost completely controlled by the pressure at 5 km. The regions of rise or fall of pressure in 24 hr. are found to move along the isobars at 5 km. in the sense of the gradient wind, keeping high pressure to the right. The changes of pressure in 3 hr. are less completely controlled by the 5-km. pressure distribution, being largely due to the motion of warm and cold air masses in the lower troposphere. The pressure distribution at 5 km. is apparently used only as a substitute for the pressure dis-

tribution in the stratosphere, but the substitution usually works well in practice when there is a moderate or large gradient of pressure at 5 km. When this pressure gradient is light the method fails, and the pressure gradient in the stratosphere, or the observed winds in the lower stratosphere or upper troposphere, must be used.

Baur's essential assumption is that there exists what may be described as a broad weather situation, which persists over a period of days, although the details of the weather of the surface layers may alter appreciably during the interval. Since the pressure gradient in the lower stratosphere, the average gradient of temperature in the troposphere, the general flow, and the steering of the regions of 24-hr. pressure changes fit together and last for some days, according to Baur, he regards it as reasonable to take them as constituting the broad weather situation. Baur finds that the mean duration of a broad weather situation is $5\frac{1}{2}$ days, and that, in Central Europe, on an average five of these situations occur in a month. The longest durations are those of east to west steering, when Central Europe is under a region of high pressure in the stratosphere, so that it is not crossed by areas of rise and fall of pressure.

It is thus seen that the problem of forecasting for 10 days ahead is essentially that of forecasting the formation and breaking down of stratospheric highs and lows. The details of the methods of forecasting have been described by Baur[97]. The practical methods are partly statistical and partly synoptic, the statistical methods using 5-day means of pressure and temperature at 26 stations as far apart as Greenland, the Azores, Algiers, Kiev and Vardo. Coefficients of correlation between data for any 10 days and the preceding 10 days were evaluated as part of the preparation of the mechanism of forecasting. But Baur has himself pointed out[98] the dangers of assuming that any evaluated coefficient of correlation can be used with machine-like precision. Thus he finds that while the coefficient of correlation between the changes of mean temperature from one year to the next at Christiansund in November to January and the same factor at Berlin in February and March was 0·73 in the years 1861 to 1889, it was −0·30 after these years up to 1918. For the numerical details of Baur's methods, reference should be made to his paper[97]. By comparison with these methods, those developed by Walker[99] for forecasting the monsoon rainfall of India appear simple, dealing as they do with purely statistical data.

§ 10. THE TROPICAL CYCLONE

Durst and Sutcliffe[100], in a discussion of the theory of the genesis of the tropical cyclone by ascent of air and consequent convergence, lay special emphasis on the importance of the vertical components of velocity and acceleration in producing the required fall of pressure by divergence in the upper air. Air forced upward with a horizontal component of velocity too great to balance the pressure gradient at higher levels will be thrown outward from the centre, and Durst and Sutcliffe regard this as one of the main factors in producing the fall of surface pressure.

Grimes[101] puts forward the view that the zones of formation of tropical cyclones in the northern hemisphere are near those portions of the equatorial front where the air coming from the southern hemisphere has a cyclonic circulation, having had an anticyclonic circulation when south of the equator. A similar view was put forward by Chapel[102] in a discussion of the air movements across the equator in relation to the development and early movement of tropical cyclones near the Isthmus of Panama.

§ 11. THE UPPER AIR

(*a*) *Ozone.* In a paper by Götz, Meetham and Dobson[103] it was shown that the amount of ozone above Arosa, measured in centimetres thickness of the gas at N.T.P. per kilometre of height, attains a maximum between 25 and 30 km. above the ground, the ratio of ozone to air reaching its maximum between 30 and 35 km. above the ground. Regener[104], using a different method, obtained substantially the same results, and it may now be accepted that these represent a close approximation to the truth. This result differs widely from those which were formerly accepted, according to which the ozone was distributed at still greater heights. A discussion of the observations, and of the effect of the ozone distribution on the distribution of temperature in the vertical[105] was given by Penndorf[106]. It can be shown that the effect of the ozone distribution is to give a region of maximum heating just below 50 km. Penndorf utters a clear warning that his numerical results are not to be taken too literally, but there is now little doubt that the observed increase of temperature with height in the stratosphere is to be explained, at least in part, by the absorption of ultra-violet radiation by ozone. The main uncertainties in the computations are due to uncertainties as to (i) the amount of water vapour in the stratosphere, (ii) the nature of the spectrum of water vapour at low pressures, and (iii) the spectral distribution of the radiation entering the stratosphere from below.

The report of the Conference on Atmospheric Ozone held at Oxford, 9–11 September 1936, published by the Royal Meteorological Society[107], contains a series of papers dealing with methods of measurement of the amount of ozone, its vertical distribution, absorption and radiation in the upper atmosphere, the relation between the horizontal distribution of ozone and weather conditions, and the general theory of the formation and dissociation of ozone. Dobson[108] states that changes of ± 25 per cent in the total amount of ozone in a vertical column of air above one place may take place within a few days, these changes being closely associated with weather conditions. Around a cyclonic depression the amount of ozone is nearly everywhere above the average for the latitude, and is particularly high just west of the centre of low pressure, i.e. above the northerly current behind the centre, where it is frequently greater than the normal amount in any latitude. Meetham[109] found that the day-to-day changes in total ozone were more closely correlated with those of potential temperature than with those of pressure or temperature, or any obvious function of these two variables.

Whipple[110] showed that the propagation of sound to great distances leads to conclusions which are in general accord with those inferred from the ozone observations, the temperature over the whole globe, winter and summer, approximating to the uniform value of 330 A.

(*b*) *Tides in the upper atmosphere.* Appleton and Weekes[111] found at the level of the Region E in the ionosphere a lunar tide many times greater than the expected lunar oscillation. The relative amplitude ($\partial p/p$) at 110 km. is 0·068, as compared with the value 0·0000115 at ground level given by Chapman[112]. The mean variation with lunar hour of the equivalent height of the region *E* has an amplitude of 1 km.

(*c*) *Atmospheric oscillations.* Pekeris[113] has shown that if the increase of temperature between 30 and 60 km. inferred by Whipple[110] is assumed, it is possible to find a temperature distribution above 60 km. such that the atmosphere has a free period of oscillation very close to 12 hr. He finds a horizontal nodal surface at 30 km., the oscillations being in opposite directions above and below this height. The same atmosphere has also another mode of free oscillation with a period of 10½ hr., but with no nodal surface. This period is in agreement with the deductions of G. I. Taylor[114] from the waves caused by the eruption of Krakatoa in 1883, these waves having a speed of propagation, in agreement with that found by Whipple[115] for the waves from the Great Siberian Meteor, corresponding to a free period of 10½ hr.

§ 12. A NEW TYPE OF INSTRUMENT, THE RADIO-METEOROGRAPH

New types of instruments have been developed in recent years in many countries for the observation of temperature and humidity in the upper air, using a radio-transmitter attached to a balloon. Such instruments have the advantage of being capable of use in all weather conditions, and of giving rapid results. A comparison of results obtained by the Harvard Radio-Meteorograph and a conventional balloon meteorograph, made by Lange[116], appears to indicate that the radio-meteorograph has reached a high standard of reliability. The Harvard instrument has been described by Lange[117].

Radio-meteorographs have also been devised in other countries: in Russia by Moltschanoff[118]; in France by Bureau[119]; in Germany by Duckert[120]; and in Finland by Vaisala[121].

§ 13. LIGHTNING AND THUNDERSTORMS

A series of papers by Schonland and various collaborators[122] has clarified our ideas as to the nature of the lightning discharge. A lightning stroke between a thundercloud and the ground starts from the cloud as a leader stroke travelling downward, the leader being a negative or cathode stream. This is followed by a second stage in which there is an upward-moving or anode stream. Each of these streamers is a conducting filament of ionized gas, which extends its length by virtue

of the ionizing processes in the strong electric field at its tip. Subsequent lightning strokes follow the path traced out by the first stroke. The total charge conveyed to the ground is negative in the great majority of cases. The cloud end is therefore generally a cathode, though it is not yet established that this holds in detail for each of the component strokes of a discharge. Norinder[123] reported that observations with a cathode-ray oscillograph in Sweden showed that a positive discharge stroke is sometimes followed by a number of negative discharges, but Schonland found no confirmation of this in his observations in South Africa, where all the strokes of a discharge were observed to be a lowering and sudden destruction of a negative cloud charge.

Simpson and Scrase[124], in an investigation of the distribution of electricity in a thundercloud, found that in general a thundercloud has a positive charge in the upper layers, and a negative charge in the lower layers within the cloud, with very frequently a region of positive charge at the base. This lower positively charged region is usually found in the neighbourhood of the active centre of the storm, being associated with heavy rain, nearly always positively charged. The region separating the upper positive charge from the lower negative charge occurs at levels where the temperature is below the freezing point, and Simpson and Scrase suggest that the generation of the positive charge depends on the presence of ice crystals, and not on the presence of water drops. The position of the lower positive charge supports the idea that it is generated by the breaking drop process described by Simpson[125].

The electricity in rain generally has been described by Scrase[126]. Three-quarters of the rain at Kew was found to be positively charged, but the charge per unit volume of rain was in general much greater on the negatively charged rain than on the positively charged rain. As a result, the total quantities of positive and negative electricity brought down by rain do not differ much, the positive charge being on the whole slightly in excess. Showery rain proved to be responsible for most of the high negative charges, continuous rain and thunderstorms being generally associated with an excess of positive charge.

REFERENCES

(1) Bennett, M. G. *Quart. J. R. Met. Soc.* **60**, 3 (1934).
(2) Owens, J. S. *Proc. Roy. Soc.* A, **110**, 738 (1926).
(3) Pick, W. H. *Quart. J. R. Met. Soc.* **57**, 238 (1931).
(4) Sohoni, V. V. and Paranjpe, M. M. *Quart. J. R. Met. Soc.* **60**, 15 (1934).
(5) Köhler, H. *Ark. Mat. Astr. Fys.* A, **24**, no. 19 (1933).
(6) Defant, A. *S.B. Akad. Wiss. Wien*, **114**, 585 (1905).
(7) Niederdorfer, E. *Met. Z.* **49**, 1 (1932); **53**, 33 (1936).
(8) Kneusel, St. *Met. Z.* **52**, 64 (1935).
(9) Bergeron, T. *Rep. Met. Assoc. U.G.G.I.* 156 (1933).
(10) Findeisen, W. *Met. Z.* **55**, 121 (1938).
(11) Bigg, W. H. *Prof. Notes Met. Off., Lond.*, no. 81 (1937).
(12) Simpson, G. C. *Prof. Notes Met. Off., Lond.*, no. 82 (1937).
(13) Noth and Polte. *Luftwissen*, Nov. (1935).

(14) Byers, R. B. *Synoptic and Aerological Meteorology*, chap. xiii.
(15) Bleeker, W. *Met. Z.* **49**, 349 (1932).
(16) Schumann, T. E. W. *Quart. J. R. Met. Soc.* **64**, 3 (1938).
(17) Bilham, E. G. and Relf, E. F. *Quart. J. R. Met. Soc.* **63**, 149 (1937).
(18) Chandra, K. *Proc. Roy. Soc.* A, **164**, 231 (1938).
(19) Brunt, D. *Quart. J. R. Met. Soc.* **63**, 277 (1937).
(20) Rayleigh. *Coll. Papers*, **6**, 432.
Jeffreys, H. *Phil. Mag.* **2**, 833 (1926); *Proc. Roy. Soc.* A, **118**, 195 (1928).
(21) Brunt, D. *Quart. J. R. Met. Soc.* **63**, 277, figs. 15 and 17 (1937).
(22) Walker, G. T. *Quart. J. R. Met. Soc.* **57**, 413 (1931).
(23) Suckstorff, G. A. *Met. Z.* **52**, 449 (1935).
(24) Poulter, R. M. *Quart. J. R. Met. Soc.* **64**, 277 (1938).
(25) Krick, J. P. *J. Aero. Sci.* **4**, no. 9 (1937).
(26) Bennett, M. G. *Quart. J. R. Met. Soc.* **56**, 1 (1930).
(27) Middleton, W. E. K. *Visibility in Meteorology* (Toronto, 1935).
(28) Wright, L. G. *Quart. J. R. Met. Soc.* **61**, 71 (1935).
(29) Bennett, M. G. *Quart. J. R. Met. Soc.* **61**, 179 (1935).
(30) Neuhoff. *Abh. K. Preuss. met. Inst.* **1**, 271 (1900).
(31) Shaw, W. N. *Manual of Meteorology*, **3**.
(32) Rossby, C.-G. *M.I.T. Papers in Phys. Oceanography and Met.* **1**, no. 3 (1932).
(33) Refsdal, A. *Met. Z.* **52**, 1 (1935).
(34) Normand, C. W. B. *Mem. Indian Met. Dep.* **23**, 1 (1921).
(35) Hewson, E. W. I, *Quart. J. R. Met. Soc.* **62**, 387 (1936); II, **63**, 7 (1937); III, **63**, 328 (1937); IV, **64**, 407 (1938).
(36) Bjerknes, J. *Geofys. Publ.* **9**, no. 9 (1931).
(37) Normand, C. W. B. *Quart. J. R. Met. Soc.* **64**, 47 (1938).
(38) Bjerknes, J. *Quart. J. R. Met. Soc.* **64**, 325 (1938).
(39) Burmeister. *Braunschweig. Verh. deuts. phys. Ges.* **15**, 612 (1913).
(40) Hettner, G. *Ann. Phys., Lpz.*, **55**, 476 (1918).
(41) Weber and Randall. *Phys. Rev.* **40**, 835 (1932).
(42) Wexler, H. *Mon. Weath. Rev.* **65**, 122 (1936).
(43) Elsasser, W. M. *Mon. Weath. Rev.* **65**, 323 (1937); **66**, 175 (1938).
(44) Dennison, D. M. *Phys. Rev.* **31**, 503 (1928).
(45) Albrecht, F. *Met. Z.* **48**, 57 (1931).
(46) Simpson, G. C. *Mem. R. Met. Soc.* **3**, no. 21 (1928).
(47) Brunt, D. *Proc. Roy. Soc.* A, **124**, 201 (1929).
(48) Möller and Mügge. *Beitr. Phys. frei. Atm.* **20**, 220 (1932).
(49) Baur and Philipps. *Beitr. Geophys.* **45**, 83 (1935).
(50) Brunt, D. *Quart. J. R. Met. Soc.* **58**, 389 (1932).
(51) Durst, C. S. *Met. Mag.* (Jan. 1932).
(52) Taylor, G. I. *Proc. Lond. Math. Soc.* **20**, 196 (1922).
(53) Sutton, O. G. *Proc. Roy. Soc.* A, **135**, 143 (1932); **146**, 701 (1934).
(54) Rossby and Montgomery. *M.I.T. Papers in Phys. Oceanography and Met.* **3**, no. 3 (1935).
(55) Mildner, P. *Beitr. Phys. frei. Atm.* **19**, 151 (1932).
(56) Taylor, G. I. *Proc. Roy. Soc.* A, **92**, 196 (1916).
(57) Sutcliffe, R. C. *Quart. J. R. Met. Soc.* **62**, 3 (1936).
(58) Durst, C. S. *Quart. J. R. Met. Soc.* **62**, 290 (1936).
(59) Sutton, O. G. *Quart. J. R. Met. Soc.* **62**, 124 (1936).
(60) Sverdrup, H. U. *Geofys. Publ.* **11**, no. 7 (1936); *Met. Z.* **53**, 10 (1936).
(61) Berg, H. *Beitr. Phys. frei. Atm.* **23**, 143 (1936).
(62) Lettau, H. *Ann. Hydrogr., Berl.*, **65**, H. 4 (1937).
(63) Powell and Griffiths. *Trans. Instn Chem. Engrs*, **13**, 175 (1935).
(64) Millar, F. G. *Can. Met. Mem.* **1**, no. 2 (1937).
(65) Lettau and Dorffel. *Ann. Hydrogr., Berl.*, **64**, 342 (1936).
(66) Elias, F. *Z. angew. Math. Mech.* 9 and 10 (1929) and (1930).

(67) GEIGER, R. *Das Klima der bodennahen Luftschichten* (Braunschweig, 1928).
(68) SCHMIDT, W. *Quart. J. R. Met. Soc.* **60**, 345 (1934).
(69) HAWKE, E. L. *Quart. J. R. Met. Soc.* **59**, 261 (1933).
(70) CORNFORD, C. E. *Quart. J. R. Met. Soc.* **64**, 553 (1938).
(71) RAMDAS. *Reports of Agricultural Met. Section, Indian Met. Dep.*
(72) HEYWOOD, G. P. *Quart. J. R. Met. Soc.* **59**, 47 (1933).
(73) SOLBERG, H. See *Physikalische Hydrodynamik*, by Bjerknes, V., Bjerknes, J., Solberg, H. and Bergeron, T.
(74) BJERKNES, J. and GODSKE, C. L. *Astrophys. Norveg.* **1**, no. 6 (1936).
(75) GOLDIE, A. H. R. *Geophys. Mem.* no. 72 (1937).
(76) BJERKNES, J. *Geofys. Publ.* **9**, no. 9 (1931); **11**, no. 4 (1931).
(77) BJERKNES, J. and PALMÉN, E. *Beitr. Phys. frei. Atm.* **21**, 53 (1933).
(78) BJERKNES, J. *Met. Z.* **54**, 462 (1937).
(79) SCHMIEDEL, K. *Veröff. Geoph. Inst. Lpz.* ser. II, **9**, H. 1 (1937).
(80) MÜGGE, R. *Met. Z.* **55**, 1 (1938).
(81) SCHERHAG, R. *Met. Z.* **51**, 129 (1934).
(82) FICKER, H. *Met. Z.* **55**, 8 (1938).
(83) MOLLWO, H. *Beitr. Phys. frei. Atm.* **23**, 199 (1936).
(84) RAETHJEN, P. *Met. Z.* **54**, 393 (1937).
(85) MÜGGE, R. *Met. Z.* **55**, 197 (1938).
(86) SCHERHAG, R. *Met. Z.* **51**, 129 (1934); *Ann. Hydrogr., Berl.*, **62**, 397 (1934).
(87) KÖPPEN, W. *Ann. Hydrogr., Berl.*, **61**, 374 (1933).
(88) THOMAS, H. *Met. Z.* **52**, 41 (1935).
(89) SCHERHAG, R. *Met. Z.* **53**, 10 (1936).
(90) BJERKNES, J. *Geofys. Publ.* **9**, no. 9.
(91) HEWSON, E. W. *Quart. J. R. Met. Soc.* **63**, 7 (1937).
(92) DOUGLAS, C. K. M. *Quart. J. R. Met. Soc.* **48**, 342 (1922).
(93) PALMÉN, E. *Met. Z.* **48**, 281 (1931).
(94) DOUGLAS, C. K. M. *Quart. J. R. Met. Soc.* **60**, 143 (1934).
(95) BAUR, F. *Met. Z.* **53**, 237 (1936).
(96) WALKER, G. T. *Quart. J. R. Met. Soc.* **63**, 471 (1937).
(97) BAUR, F. *Bull. Amer. Met. Soc.* **17**, 148 and 252 (1936).
(98) BAUR, F. *Met. Z.* **54**, 188 (1937).
(99) WALKER, G. T. *Mem. R. Met. Soc.* **3**, no. 24 (1930); **4**, no. 36 (1932); **4**, no. 39 (1937).
(100) DURST, C. S. and SUTCLIFFE, R. C. *Quart. J. R. Met. Soc.* **64**, 75 (1938).
(101) GRIMES, A. *Mem. Malayan Met. Serv.* no. 2 (1937).
(102) CHAPEL, L. T. *Mon. Weath. Rev.* **62**, 433 (1934).
(103) GÖTZ, MEETHAM and DOBSON. *Proc. Roy. Soc.* A, **145**, 146 (1934).
(104) REGENER, E. *Beitr. Phys. frei. Atm.* **22**, 249 (1935).
(105) See JAUMOTTE. *Beitr. Geophys.* **50**, 403 (1937).
(106) PENNDORF, R. *Veröff. Geoph. Inst. Lpz.* ser. II, **8**, H. 4 (1936).
(107) *Quart. J. R. Met. Soc.* **62**, Supplement (1936).
(108) DOBSON, G. M. B. *Quart. J. R. Met. Soc.* 62, Supplement, p. 52 (1936).
(109) MEETHAM, A. R. *Quart. J. R. Met. Soc.* **62**, Supplement, p. 63 (1936); also *Quart. J. R. Met. Soc.* **63**, 289 (1937).
(110) WHIPPLE, F. J. W. *Quart. J. R. Met. Soc.* 62, Supplement, p. 19 (1936); also *Quart. J. R. Met. Soc.* **61**, 285 (1935).
111) APPLETON and WEEKES. *Nature, Lond.*, **142**, 71 (1938).
(112) CHAPMAN, S. *Quart. J. R. Met. Soc.* **44**, 271 (1918).
(113) PEKERIS, C. L. *Proc. Roy. Soc.* A, **158**, 650 (1937).
(114) TAYLOR, G. I. *Mem. R. Met. Soc.* **4**, no. 35 (1932).
(115) WHIPPLE, F. J. W. *Quart. J. R. Met. Soc.* **60**, 510 (1934).
(116) LANGE, K. O. *Beitr. Phys. frei. Atm.* **24**, 243 (1938).
(117) LANGE, K. O. *Bull. Amer. Met. Soc.* May 1936, March 1937.
(118) MOLTSCHANOFF, P. A. *Met. Hydrol.* no. 2, 30 (1936).

(119) Bureau, R. *Ondes Electr.* **14**, 10 and 87 (1935).
(120) Duckert, P. *Beitr. Phys. frei. Atm.* **20**, 303 (1933).
(121) Vaisala, V. *Helsingfors Comm. phys. math.* **8**, no. 4, 12 (1936); **9**, no. 9, 78 (1937).
(122) Schonland and others. *Proc. Roy. Soc.* A, **143**, 654 (1934); **152**, 595 (1935); **162**, 175 (1937); **164**, 132 (1938).
(123) Norinder, H. *J. Franklin Inst.* **225**, 69 (1936).
(124) Simpson, G. C. and Scrase, F. G. *Proc. Roy. Soc.* A, **161**, 309 (1937).
(125) Simpson, G. C. *Proc. Roy. Soc.* A, **114**, 376 (1927).
(126) Scrase, F. G. *Geophys. Mem.* no. 75 (1938).

SOUND

By E. G. RICHARDSON, B.A., Ph.D., D.Sc.

§ 1. FREE PROPAGATION: SUPERSONICS

Supersonic dispersion in gases and liquids still attracts the majority of workers in this field. We shall first deal with the published work in 1937 which forms an extension of the matters which were treated in a special *Report* in Volume 4.

Taking first the work on gases, we find that Hodge and Hubbard[1,2] have extended the earlier measurements on the change of supersonic velocity with pressure up to 100 atm. using a Pierce interferometer and five frequencies from 88 to 500 kc./sec. The chamber was constructed out of a block of steel with a cap which could be removed to replace the quartz oscillator. The metal to metal joint which permitted the apparatus to be sealed was made by cutting a circular V-shaped ring into a soft copper gasket with sufficient force to cause the copper to flow. The force to do this was applied by screwing down two bolts with a wrench, making a gas tight seal even at 100 atm. The pressure was measured on a Bourdon gauge. For the results: a steady rise was observed in hydrogen, helium, oxygen and nitrogen, but a fall with increasing pressure was apparent in carbon dioxide. No evidence of the usual S-shaped curve, observed in the case of carbon dioxide by earlier workers, appears in the results, but this is possibly because pressures below atmospheric were not applied. In conjunction with known values of compressibility, the ratio of specific heats for these gases at various pressures is then deduced. Eucken continues his work, this time in collaboration with Nümann[3], on gases at high temperatures. Both absorption and velocity in the dispersive gases, carbon dioxide and nitrous oxide, are measured up to 400° C., with attendant variations of pressure from normal up to 4 atm. and a single frequency of 200 kc./sec. The reflector is kept at a fixed distance from the source while the pressure is varied, and it is assumed that modifications in the anode current are then entirely due to changes in the absorptive capacity of the gas. (In the writer's experience such an assumption is unjustified unless supported by independent experimental evidence, since a change of gas pressure alters the load on the crystal and with it the driving current). Warner[4] makes velocity measurements only in carbon dioxide, nitrous oxide and sulphur dioxide over a smaller range of temperature and at frequencies between 40 and 100 kc./sec. His results in general confirm and extend those of Penman (cf. volume 3, p. 107) in that rise of temperature traces out the S curve from top to bottom in the former gases, whereas sulphur dioxide shows constancy of velocity within the temperature and frequency range chosen. He plans to increase the range of temperature and to pursue the frequency down to audible pitch. Itterbeek and Mariëns[5] have worked in oxygen, nitrogen, hydrogen and carbon dioxide at low temperatures, expressing both velocity and attenuation coefficients as functions of pressure (down to one-tenth of an atmosphere) for a frequency of 300 kc./sec. At

90° K. the velocity falls steadily in the atmospheric gases as the pressure rises, while at all temperatures the absorption coefficient drops steeply at first but more gradually as normal pressure is approached. On the hypothesis that the absorption is a function of the mean relaxation time for the oscillations of the molecules, they expected that if a paramagnetic gas like oxygen were placed in a magnetic field it might show altered values for the propagation constants of supersonic radiation. No change of velocity within estimated experimental error was observed, but they infer a change in absorption. It must be remembered, however, that with their apparatus the absorption cannot be measured to the same accuracy as the wavelength and they seem themselves loath to attach too much significance to this inference. Examining results for mixtures of carbon dioxide with small adulterations of foreign gases, Metter[6] comes to the conclusion that change of dipole moment—which has been suggested as a possible cause of the dispersion of supersonics—does not run parallel with the supersonic observations, but that a more consistent parallel is to be observed with the energy-balance of the possible chemical reactions between the two sets of molecules.

Turning now to observations in liquids, which are mostly of abnormal attenuation, we find that Bazulin[7] is continuing his work on the effect of temperature in organic liquids. Between 20 and 80° C. he finds that the absorption coefficient rises linearly with temperature in benzene at 11,500 kc./sec., the increase being about 50 per cent. Apart from the fact that the values are far in excess of those of Kirchhoff's theory, purely viscous absorption should of course decrease as the temperature goes up. In carbon tetrachloride, on the other hand, he notes constancy of attenuation until the boiling point is approached. Between 60 and 80° C. the results are much affected by pre-treatment of the liquid. If this has been well degassed by prolonged boiling, little change in absorption is observed as cooling through this temperature range takes place. In the reverse direction, fresh liquid showed a rise of attenuation, evidently due to bubbles of dissolved gas acting as scattering centres. Korolev[8] confirms the benzene results by taking *schlieren* photographs with light cast athwart the supersonic field. Shadows are produced on the exposed plate from which the intensities of the diffracting planes (compressions in the medium) at various distances from the source are calculated. Bergmann and Goehlich[9] have devised a simple arrangement for studying this effect which resembles that of Nomoto[10] published in 1937 and since amplified by him. The quartz is at the bottom of a trough containing the liquid. A few centimetres to one side is fixed a 4-W. lamp, having a straight horizontal filament about 12 mm. long, mounted in a lantern with an iris diaphragm on the trough side. The filament then lies parallel to the wave front and casts an image of the wave system on a screen on the far side of the trough. For visual demonstration, a reflector is adjusted at the top of the tank to give a system of supersonic standing waves. Nodal planes then appear as dark lines on the screen. (This arrangement makes a good lecture demonstration, if the power of the lamp and the magnification be increased.) Very similar is the apparatus of Claeys, Errera and Sack[11], save that they let the supersonic energy exert pressure on a sphere 2 cm. diameter, the liquid being contained

in a tube 12 cm. wide. Their results are of the same order as Biquard's (cf. volume 4, p. 69).

Schaafs[12] uses the familiar optical diffraction method to study the velocity in a number of pure liquids and mixtures at 16,000 kc./sec. In most cases the relation between velocity and concentration of the adjunct is a linear one, sufficiently so to enable extrapolation to the velocity for the pure solute to be made. After this similar measurements were made with solid substances forming one of the constituents, in order to see to what extent extrapolation to the velocity in the *solid* material could be attempted. Although data for confirming the extrapolated values are rather meagre, they are sufficient to indicate that this will prove a useful method in physical chemistry for deducing data about substances in special physical states which could not otherwise be obtained. Wyss[13] uses a similar method for the *stationary* waves formed between the quartz and the reflector, passing a narrow beam of light across the field of the supersonic "grating" and letting it fall on a revolving film camera. As the reflector is moved, maxima and minima in the diffracted beam of light appear according as a node or antinode is formed at the place of observation. The method is thus a combination of the Pierce with the Brillouin dispositions. He uses the apparatus to measure supersonic velocities in a number of organic liquids and in sugar solutions of varying concentration (in which a linear change of velocity is also evident). Bär[14] has pushed absorption measurements in liquids up to the enormous frequency of 80 Mc./sec. using the optical method. On the Kirchhoff theory the absorption coefficient divided by the square of the frequency is about 10^{-16} for most liquids. Bär finds values of four times this (confirming the previous results at lower supersonic frequencies in water and a number of organic liquids); but in carbon disulphide the absorption is a thousand times greater than theory, while mercury gives very nearly the theoretical value (at 50 Mc./sec.). This last point is interesting as mercury had not previously been used in this method. Of course, as mercury is optically opaque, observations could not be made in it directly, but a water layer was put above it, the energy getting through various depths of mercury to the water being recorded in the optical system. A method used by Otpushtshennikov[15] which has several advantages over other methods for absorption in liquids and solids is not new, at any rate in principle, consisting as it does of a high power quartz oscillator facing a quartz receiver of equal frequency. The author assumes that the radiation picked up by the receiver falls off exponentially with distance from the source, and the correction (due to the stationary waves formed at certain separations) must be negligible. This does not tally with the experience of other users of this method and probably accounts for his results in liquids being in disaccord with the majority. He has also been able to apply the method to metallic media but says with more becoming diffidence "the absorption has not yet been evaluated because of the uncertainty as to what extent stationary waves form in the metal and how they influence the result of the absorption measurement". Fox[16] has attacked the problem from a different aspect. He considers the impedance theory of the quartz oscillator when coupled to a fluid column bounded at the far end by a rigid reflector,

while the length of the column is varied, and shows how to deduce both the attenuation constant of the liquid and the reflection coefficient at the solid boundary in terms of indications in the driving circuit. He tests his theory on a water column at 3 and 8 Mc./sec. using monel and brass reflectors. His constant for water is a little lower than that of earlier observers, but he points out that he used far less power to drive his oscillator. One practical point which he notes is worthy to be stressed. When a quartz is loaded with a liquid column the crevasses in the anode current as the length is varied are much wider than with a gaseous load. Consequently, if the crystal tends to vibrate on two nearby frequencies, errors in wavelength due to this cause are more likely to pass unnoticed.

Dispersion of the velocity in liquids does not yet seem established outside the limits of experimental error though Rao[17] notices an isolated instance. As between carbon tetrachloride and acetone he found a higher velocity in the former at 5×10^9 c./sec., in the latter at 7×10^6 c./sec.

We referred in our special *Report* last year to the concept of a sound scattering medium as being one which is speckled or, as Lucas[18] aptly puts it, *élastiquement troublé*. This author takes up the well-known theory of Stokes for the attenuation of viscosity and applies it to a medium which contains fluctuations of density. It is suggested that thermal agitation may give rise in pure liquids to local variations in both density and compressibility, sufficient to diffuse a system of plane waves from a forthright path. This process—as the writer has already pointed out—should be analogous to that by which light is scattered from a colloidal solution (Mie effect). This naturally brings us to the discussion of further work on fluids containing actual diffusing particles. Brandt[19], Freund and Hiedemann[20] have elaborated the work that they described at the Faraday Society Discussion on Dust and Smoke in 1936. As far as the calculation of the relative amplitudes of solids and fluid particles in vibration are concerned there is little that is new, but they have extended the theory to evaluate the attenuation of the vibration, due to viscous drag in the relative motion of the two phases; treating the solid particles as rigid. For particles between 0·5 and 2μ. in diameter the absorption rises in the supersonic region (the more steeply the smaller the particles, but starting to rise at a higher value of the frequency), since the larger particles are unable to follow the gaseous motion even at moderate frequencies. The authors go on to discuss the coagulation of particles under the action of their relative motion in a sound-irradiated medium and illustrate their remarks by some experiments in which the rate of coagulation of a tobacco smoke is followed in terms of the light scattered by the suspension.

Landau and Rumer[21] have tackled the theory of supersonic absorption in solids, following a line of reasoning somewhat similar to that of Lucas. The absorption is supposed to arise from interaction between supersonic quanta and Debye heat quanta, of which the wavelengths must be large compared with those of the sound; this implies low temperature. The ideas are developed as follows. A plane sound wave represented by the vector f and pulsatance w collides with a wave of thermal agitation (f_1 and w_1). Within classical bounds no interchange of energy can take place between them, otherwise the principle of superposition could not hold.

But if cubic terms are included in the expression for the energy density—which implies distortion—then transitions take place by which energy and momentum are carried over from the sound wave into the Debye wave, so that the outcome is a new wave (f_2 and w_2). Then $f_2 = f + f_1$ and $w_2 = w + w_1$, but as with most of these quanta exchanges $\lambda_2 \rightarrow \lambda + \lambda_1$. Hence the thermal waves in the solid must have a smaller velocity than that of the sound for this process to take place, and so absorption due to this cause is most likely to occur as between *transverse* sound waves and *longitudinal* thermal waves.

We may now pass on to those supersonic experiments in which the emphasis is rather on the behaviour of the source and its reaction on the surroundings than on the properties of the irradiated medium. Rzevkin[22] has examined the field in vaseline oil using the *schlieren* method under the stroboscopic illumination of a Kerr cell. This made it possible to examine the propagation of progressive waves in the oil. He gets complicated patterns which he ascribes to interference between the longitudinal waves sent out by the quartz and others of lower frequency due to its flexural vibrations. Very similar patterns are obtained if the supersonic radiation is sent through a grating. Hiedemann and Hoesch[23] have also probed the field in front of the quartz, using their own version of the optical *schlieren* effect and have shown that the amplitude varies considerably from place to place. The field shows a fine structure of many criss-cross rays. Continuing the same research, Schreuer and Osterhammel[24] have placed a plane semi-permeable reflector at a slight angle to the wave front in a supersonic stationary wave pattern, so as to photograph the interference effects between the stationary and the progressive reflected waves. (The latter are absorbed at the walls of the vessel and so do not return to the field of view.) Similar patterns for two sets of crossing progressive waves are shown. Next Hiedemann and Osterhammel[25] exhibit some very beautiful photographs of the interference field in front of quartz oscillators working under conditions such that there is a progressively increasing ratio between the face width d (which acts as the diffracting aperture) and the wavelength λ. By permission of Dr Hiedemann, the two extreme cases are reproduced here, figure 1, (*a*) when $d = 1{\cdot}3\lambda$ and (*b*) $d = 7{\cdot}5\lambda$. Apart from interference lines the brightness of the field shows to what extent the energy is being concentrated in a beam. Finally Hiedemann and Schreuer[26] draw attention to the secondary interference lines observed when light is incident on a supersonic grating. These arise in the same way as those in light diffraction at an ordinary grating, i.e. by interference between the diffracted rays from neighbouring elements. Bär[27] has also carefully examined the diffraction patterns of supersonic radiation in a number of cases analogous to those used in light, i.e. passage through an aperture and round an obstacle, Lloyd's fringes etc., in order to see from the clarity of the fringes how closely the vibration of the piezo-electric oscillator corresponds with that of the ideal piston radiating through a hole in a wall. In the case of the quartz examined, the fundamental (440 kc./sec.) and lower partial vibrations were far from uniform in phase and amplitude over the face of the quartz, but in the range 5000 to 10000 kc./sec. the vibrations were very nearly of the piston type. (Some departure from the ideal type is apparent in Hiedemann and Osterhammel's

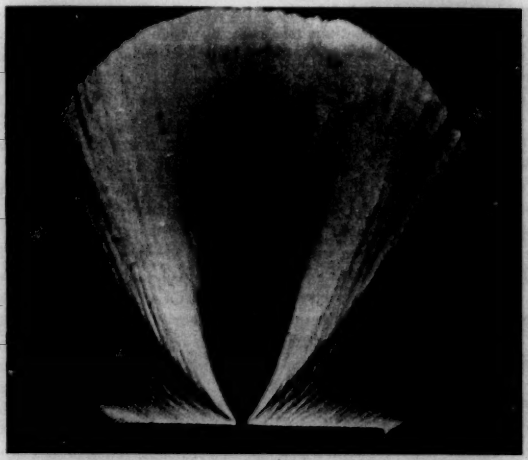


Figure 1 (*a*).

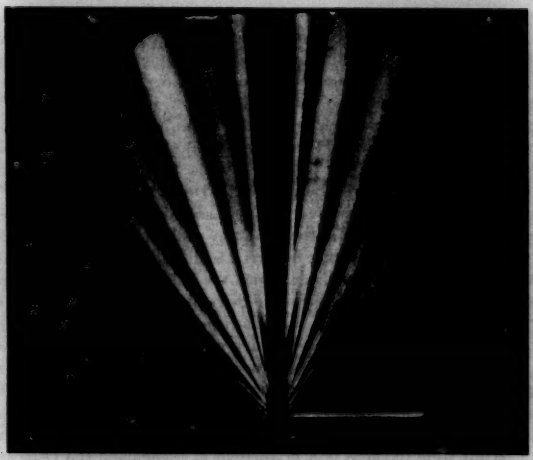


Figure 1 (*b*).

photographs, cf. figure 1). Baumgardt[28] has employed his apparatus for the radiation pressure in supersonic waves passing through liquids to measure the output of energy from the quartz in its overtone vibrations. Taking the lower surface of the quartz as immobile against its electrode, while the other in contact with the liquid has amplitude a, the supersonic power at pulsatance w is given by

$$P=\tfrac{1}{2}w^2\rho Va^2 \qquad \text{.....(1)}$$

and the amplitude a by

$$a=\frac{\rho_1 V_1}{\rho_0 V_0}\cdot\frac{2\delta E}{(2n+1)\pi}, \qquad \text{.....(2)}$$

where ρ_0, ρ_1 and V_0, V_1 are the density and velocity in the liquid and quartz respectively; δ the piezo-electric modulus and E the applied potential; $n=0, 1, 2,$ etc. and denotes the order of the partial vibration. For each mode, as the driving frequency is varied, he measures P (by the pressure of the radiation) and E (by an electrometer) and compares the practical output of power with that calculated from (1) and (2). The agreement is good at the fundamental, but in fact the power falls off in the higher overtones as the table shows:

Mode	Liquid	Power (w./cm.^{3} v.) $\times 10^7$	
		actual	calc.
1	Water	3·6	3·8
3	Benzene	0·37	0·62
3	Toluene	0·40	0·62
5	Benzene	0·29	0·62
5	Toluene	0·31	0·62

The number of direct researches on high frequency sources seem to be fewer than in earlier years. Hiltscher[29], however, has conducted an interesting research into the possibility of exciting piezo-electric oscillations in rochelle salt crystals. In these deformation can only occur in a plane perpendicular to the applied field so that a plate of rochelle salt, in contra-distinction to quartz, can only be excited in the thickness mode. The moduli are less than one-tenth of those in quartz with a corresponding reduction in output to about 1 per cent (cf. eqns. (1) and (2) above). Further the high dielectric constant results in loss of potential when the crystal is in air. Hiltscher therefore stuck on electrodes of aluminium leaf with canada balsam. The modes of vibration of the crystal, coupled to a variable oscillator, were observed by the discharge light which is seen at the nodes in a vacuum (Giebe-Scheibe effect). The hope of replacing expensive quartz by readily grown rochelle salt crystals as practical oscillators seems unlikely to materialize, as—apart from their poor efficiency—it would mean building up a complicated mosaic of crystals to produce a radiating face of reasonable size.

Ostrovsky[30] has experimented with powerful magneto-strictive oscillators formed of nickel tubes 10 mm. in diameter and with very thin walls (0·1 to 0·2 mm. thick), a disk of the same diameter being copper-soldered to one end to act as radiator. The tube had its mid-point fixed in a rubber stopper, inserted into a bottomless glass bottle, which contained the liquid in which emulsifying experiments of the usual

type were carried out. The voltage of the generating circuit was as high as 4000. A nearby magnet produced a permanent magnetization in the rod of 600 gauss, this having been found to be the optimal value.

Accounts of special effects of supersonic radiation are—not with regret—becoming rarer, but we should report the efforts of Brohult[31] to split the haemocyanin molecule, if only to commend the probity with which a control experiment was performed to see how far the effects could be reproduced by purely thermal means! The changes were, moreover, recorded by physical and not chemical means. The rate of sedimentation of the original haemocyanin (which has a molecular weight of $6{\cdot}74 \times 10^6$) was determined in the centrifuge. After irradiation at 250 kc./sec. some of the molecules had been broken into $\frac{1}{2}$ or $\frac{1}{8}$ fragments during which time the temperature had risen to 40° c. As it required heat treatment to 60° c. before any change of molecular weight—due to association—could be induced, it was concluded that here was a genuine supersonic splitting of the molecule.

Such is the preponderance of supersonic activity in this field, that there remain only four researches on the velocity of sound at audible pitch. Tyler[32] gives a description of an instructive set of experiments in which columns of air are driven by an electrically maintained telephone source, the impedance of the gaseous load as it is changed in size or shape being derived either from its reaction on the maintaining circuit or by its action on a distant telephone, acting as receiver. A circuit is given for driving the sending system through an oscillating neon tube over the frequency range 200 to 4000 c./sec. An Osglim lamp coupled to the circuit was made to illuminate a stroboscopic disk for accurate frequency determination. The current through the telephone was passed to a rectifier (phototronic cell) and sensitive galvanometer, whose deflection was observed as the acoustic configuration of the surroundings was varied. From the point of view of wavelength determinations, using variations in the length of the column between the source and a reflector, the neon tube driving circuit has the advantage of initiating a saw-tooth wave-form of energizing current which leads to sharp reaction peaks whose separation is readily measured. The peaks also diminish in size with distance from the source. From this it should be possible to calculate the attenuation coefficient as well as the propagation constant of the column. Similar measurements in single columns were made with a second telephone as receiver, but an interesting modification involving the principle of the Quincke tube was also carried out. The reader is no doubt familiar with the trombone slide by which the length of one branch of the Quincke tube is varied as the alternative path is kept fixed. Tyler attaches the output telephone to one end of the main tube of the system, the sound passes either by the short connexion or by the longer by-pass to the other end of the main tube where the receiver is to be found. Moving the slide causes fluctuations in the received energy, the maxima occurring at slide positions a wavelength apart. Again there is evidence —by progressive diminution in size of maxima as the slide is pulled out—of damping in the longer by-pass, particularly at the higher frequencies. Free air results were obtained by putting source and receiver in a box heavily lagged with cotton wool and also by setting up an acoustical analogue of the Lloyd's mirror experiment, the

requisite variation in path difference to produce maxima and minima at the receiver being ensured by slow rotation of the mirror—a glass plate about 1 metre square. (The turntable holding the mirror was carefully lagged with cotton wool.) Finally the sensitivity of the various driving circuits is examined by observing the change in reaction on them as the length of the tube is slightly varied about that proper to the natural frequency of the source. Salceanu[33] also uses a resonance tube method to measure the velocity of sound (at 1500 c./sec.) in water and other liquids, intercalating a tube factor, due to Bungetzianu, which allows for the relative width and compressibility of containing tube and fluid. (Its derivation is not given, but for most of the liquids employed in glass tubes, the correction amounts to 2 or 3 per cent; exceptionally aniline and glycerine require the measured velocities to be raised by 10 and 15 per cent respectively).

A problem suggesting itself to Trimmer[34] by the silencers of motor cars is that of the propagation of sounds along a conduit in which the gas is moving with uniform velocity. He shows that, under the conditions which apply in practice, if a small particle velocity is superposed on a large drift velocity U, a velocity-potential of the type

$$\phi = f_1(x - c_1 t) + f_2(x + c_2 t)$$

may be postulated in which $c_1 = c + U$; $c_2 = c - U$ and c is the velocity of sound in still air. Proceeding on this assumption he exhibits the impedance looking into an open pipe of length L through which the medium is flowing with a velocity U for various values of kL_1, where $k = w/c$. One effect of the drift is a flattening of the resonant impedance peaks, while the effective wavelength in the pipe is reduced no matter which way the medium is flowing. The resonant lengths are modified, in fact, by a factor $1 - U^2/c^2$, which recalls the Fitzgerald correction in a somewhat similar phenomenon.

Dahl and Denk[35] have made a study of the propagation of sound in a gaseous medium whose motion is of more complex character, although the conditions of their experiments perhaps approach more closely those in exhaust silencers than the rather special case which Trimmer's theory covers. In some instances of the propagation of sound from fog horns over the sea they noted marked variations in time of the intensity of sound recorded by a condenser microphone at a given point in the sound field. Over a period of 3 sec., during which the output of the transmitter remained quite constant, variations of intensity at the source amounted to 10 db. This is ascribed to the turbulence in the air. Rolling and whirling air bodies distributed irregularly through the structure act as scattering centres, particularly important for a sound having a wavelength of the same order as the diameter of the air pockets. They suggest that turbulence in the atmosphere might be studied by such acoustical means.

§ 2. VIBRATING SYSTEMS: SOLIDS

The Fitzgerald contraction in the length of a body moving in the direction of its length, to which we have referred in § 1 in relation to another phenomenon, postulates a contraction of 5 parts in 10^9 in the length of a body when turned from

a direction at right angles to the earth's path round its orbit to a direction along that path. This should produce a corresponding change in the resonant frequency of such a body when in longitudinal vibration (*vide supra*). (It will be remembered that in the famous Michelson-Morley experiment the failure to observe a change in the apparent velocity of light—due to ether drift—as the apparatus was rotated led to the assumption that a compensating change of length of the apparatus had taken place during the rotation.) The National Physical Laboratory has now an elinvar rod and quartz oscillator which can be relied on to maintain a frequency to one part in 10^9 for an hour or more. An experiment suggested by Wood and carried out by Tomlinson and Essen[36] was accordingly made to see whether the frequency changed by a measurable amount as the oscillator was rotated through a right angle. Actually the result was nil, but whether this means that there is no such thing as ether drift or that the expected change of length is compensated by other factors we must leave to the relativists to settle. All that falls in our province is to note the elaborate means adopted to time the period of oscillation within the expected maximum error (2 in 10^{11}) due to inaccuracies of levelling; for an *ad hoc* experiment showed this to be the prime cause of periodic variations in frequency as the table holding the quartz crystal and its amplifier was turned. Such changes of frequency were measured by the beats between the moved and a fixed quartz oscillator, the beats being recorded on a chronograph of high precision. Of course, the temperature and pressure in the glass vessels containing the oscillators had to be very carefully controlled.

The *vibrating string* provides a number of interesting papers of application to the violin and the pianoforte. Backhaus[37], for whom the violin seems to provide an unlimited supply of physical problems, uses a continuous bow something like a band-saw which excites—in the first instance—a string stretched over bridges mounted on a metal plate. He finds how the pressure of the bow on the string affects the duration of the starting noises characteristic of the sound of the bowed string. This time was a minimum for a force of 26 g. on this skeleton violin. The second part of his work deals with the coupling between the string and the belly of the violin. Beats characteristic of a system with two degrees of freedom are exhibited by the wave-form of the sound of the instrument as picked up on a microphone. A close coupling between the natural vibrations of the string and the fixed vibrations of the wood is characteristic of a good violin; in other words, for these the material acts as a sound board without those marked resonances such as—in extreme cases—give rise to a wolf note. Meinel[38] with a similar bowing machine is interested principally in the effect of the *thickness* of the wood on the tone characteristics. If this is overdone, the spectrum is wanting in the lower frequencies (below 500 c./sec.); very thin wood exaggerates the low at the expense of the high components and the tone lacks brilliance. Putting on the varnish—a process of which an unscientific Cremona fetish has exaggerated the importance—made little difference, but moving the sound post redistributed the resonance peaks besides altering the efficiency of the apparatus, considered as a sound radiator. (The latter fact has been established before.) Some special shapes of violin (trapezoid among others) failed to improve

on the tone of the conventional shape and would be—one would imagine—awkward to play with the manually operated bow. Minnaert and Vlam[39] have endeavoured to establish the role of the bridge in coupling vibrations of string and belly in the violin. The bridge, we should remark, is underpinned by the sound post which by joining the back and front board just behind one of the feet of the bridge transmits the energy directly to the back and prevents the framework from collapsing under the pressure of the bridge. The vibrations of the bridge were observed by a beam of light reflected from a galvanometer mirror stuck on with wax. Both flexural vibrations (both sides moving together in the direction of the string) and torsional vibrations (one side moving towards the player while the other moves away) were noted, as well as the simple pivoting to and fro without bending, as it were of a rigid body. All these were strongly dependent on the frequency of the note excited, but it seemed to be immaterial which string was bowed. Each of the three modes has a favoured pitch which was altered when a load in the form of a mute was put on the bridge, but not by change in the resonance of the air cavity—accomplished here by the ingenious means of passing coal gas through the interior. Saunders[40] at Harvard has carried out an independent investigation which to some extent overlaps the ones just described. He prefers hand bowing as it gives similar spectra to the mechanical bow and because musicians, not unnaturally, are inclined to discount the results of such unconventional treatment. He considers that changes in bowing pressure alter the tone quality (presumably of the steady note) very little, change in bowing speed a little more, distance of bowing point (from the bridge) most of all. But the variations were not so great, if a medium distance were maintained, to prevent an average spectrum for each fundamental frequency being plotted. "Every peak on such a response curve is the haunt of a wolf-note which appears whenever the load on the bow is insufficient to 'kill' it." The kill results from the application of sufficient energy to wipe out the transient beats between vibrations of string and belly when left to themselves. He has made comparative tests of a number of instruments including an aluminium copy of a Stradivarius, but confesses that much work is needed before the physicist can devise tests which will tell a good violin from a bad one. From all this it seems that our university students would benefit from a course on the physical problems presented by a violin.

Dealing with the coupling between strings and soundboard on the pianoforte, Vierling[41] finds that the coupling between an individual string and the board on which it is stretched is less important for the damping than that between the two or three strings which are simultaneously struck by an individual hammer and which are, of course, nominally though never quite actually, in unison. From the point of view of tone quality, it is not desirable that this trichord should be exactly in tune, but the beating between them resulting from the mistuning, though it brightens the tone, enhances the rate of attenuation of intensity. In another paper Kock[42] by a novel analogy treats the struck string as a problem in electrical transmission. The string becomes a line into which an electric impulse is applied at a certain point and time through an inductance. This denotes the impact of the hammer. Current waves travel in both directions and are reflected at the ends with reversal of phase;

on returning, if the switch is still closed, they will tend to restore energy to the inductance. Therefore if the switch is opened just before the reflection from the far end returns, maximum transfer of energy to the line will have taken place, while the earlier arriving reflection from the near end helps to reverse the voltage on the inductance, i.e. reverse the direction of motion of the hammer. Acoustically speaking, this means that to a given hammer/string mass ratio there corresponds a best striking point along the string. The theoretical deductions are compared with some earlier experiments by George and Beckett.

Lehmann[43] has constructed a novel variable tone-generator for the low frequency range (50 to 150 c./sec.). A hollow tube of silver steel, *F*, figure 2, is held at its lower end in a clamp *E*, while being loaded at its upper end by an aluminium tube *R*, containing mercury whose mass, *S*, can be varied by displacing that in *G* by means

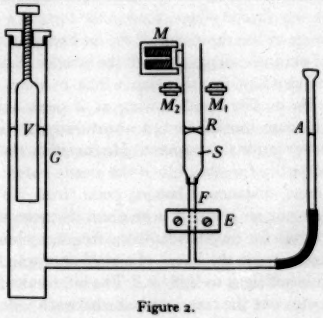


Figure 2.

of the stopper *V*. Actually the hollow reed *F* is only a few mm. long and forms the elastic component of the system. This allows the mass to be concentrated in *S*, the column of which is generally much longer than *F*. This separation of mass and elastic components allows of a frequency range over more than an octave. Under these conditions the frequency is inversely as the $\frac{3}{2}$ power of the length of the column —read off in *A*—and gives high sensitivity of the frequency to adjustments of the column. The electro-magnet M_1 maintains the oscillations in the usual valve circuit while the larger inductor *M* serves as a drive for any electro-acoustical apparatus to which it is connected. Changes of frequency with working conditions, other than by adjustment of the mercury column, are said to be very small.

Dust figures on vibrating plates continue to be produced in great variety. Pavlik[44, 45] has solved the equations of motion under various boundary conditions and obtained theoretical confirmation for the Chladni figures which he gets on uniform plates with free edges in the form of (*a*) rectangles, (*b*) parallelograms and (*c*) ellipses. The last two types appear to be new to this difficult line of theoretical research. Mary Waller[46] has used her well-known method of contact with solid carbon dioxide to

produce a number of beautiful nodal patterns on various metallic objects; some simple ones like uniform bars, others which must surely defy analysis, such as spanners. The problem of detecting laminations in welded plates in hidden locations which might give rise to serious weaknesses has led Hayes[47] of the U.S. Naval Research Laboratory to try a method based on the irregularity of the figures formed on the plate under electro-magnetic maintenance. The distortion of the pattern produced by soldering a lug under the plate to serve as an artificial defect is shown. It is deduced that lack of symmetry in a simple easily recognized pattern is a sure sign of a lamination defect, although it does not always show exactly where it is. Often a large portion of the field will be distorted by an isolated defect.

The only work on *bells* coming within our period is a short article by Taber Jones[48] who reiterates his conviction that the strike note of bells is not a difference tone. The subject is referred to again later in this article under Singing Propellers (§6).

§3. VIBRATING SYSTEMS: FLUIDS

The correlation between the frequency and intensity of the eddy production behind a solid body immersed in a fluid and the resistance which the body experiences goes back to the classical work of Kármán and Rubach in 1912. Since that time, on account of its importance in aeronautics, these aeolian vortices have acquired a formidable bibliography. Lehnert[49] has made this the subject of an inaugural dissertation in which he describes repetitions of many of the standard experiments on the subject. The resistance of a body of diameter D, moving in a fluid of density ρ with velocity U, being given by the well-known formula

$$W = k_w . \rho U^2 D;$$

Kármán and Rubach expressed the resistance coefficient in the form

$$k_w = \left[0{\cdot}799 \frac{u}{U} - 0{\cdot}323 \left(\frac{u}{U}\right)^2 \right] \frac{l}{D},$$

where u is the velocity and l the wavelength of the vortex system whose frequency f is accordingly $U-u/l$. Changes in the scale of the flow pattern, in the form of changes in u/U and of l/D are accordingly reflected in both k_w and the non-dimensional Strouhal number U/fD. The author groups the results of a number of his predecessors on a graph in which these quantities are plotted against the Reynolds number of the flow (for a circular cylinder) and amplifies this with some new results of his own (for both cylindrical wires and plates). The frequency was obtained by listening to the singing of the wire (in aeolian harp fashion) when exposed in a wind channel, picked up by microphone when too feeble to be heard by ear; while the wavelength was obtained by a pair of Pitot tubes in tandem behind the plate, connected to the ear through a T-piece. Minimum of sound occurred when the heads of the tubes were one wavelength apart. In each case the resistance of the body was measured on the usual force balance. He finds, in agreement with some earlier measurements of Tyler, that the spacing, and with it the frequency of the vortices behind the plate, changes as they pass down the stream.

Brown[50], whose work on sensitive flames we have noted in previous Reports, has now turned his attention to the vortices causing edge tones. As is well known, when a fluid issues from a linear slit to strike a wedge in its path, pendular motion of the fluid is set up, which on the one hand bears a close relationship to the aeolian vortex system, since there are alternate and equispaced convolutions of the edges of the emergent jet, and on the other hand has a certain similarity to the forking characteristic of the sensitive flame. Brown has brought out the latter affinity very well in a series of excellent smoke photographs and has further repeated the standard experiments on the variation of frequency with velocity (U) and distance from slit to edge, h. The criterion which defines the regime now is U/fh. He denies that l the inter-spacing is approximately a sub-multiple of h, the perhaps too simple assumption that lies at the basis of most edge tone theories. Over a considerable range of h, however, two simultaneous frequencies are produced, and he proposes the following empirical formula

$$f = 0{\cdot}466j\,(U - 40)\,(1/h - 0{\cdot}07), \qquad \text{.....(3)}$$

where $j = 1, 2{\cdot}3, 3{\cdot}8, 5{\cdot}4$, for the first four stages of subdivision of the slit-edge space into one or more wavelengths. (For comparison we may write the more usual but approximate formula as

$$f = 0{\cdot}5jU/h, \text{ with } j = 1, 2, 3, 4, \text{ etc.})$$

In the second part of his work he examines the more plausible theories which have been put forward to explain the mechanism of edge tone production, viz. (1) the Kármán vortex-street theory, which explains the disposition of the vortices in terms of the stability of the system, but does not properly account for the intervention of h in the expression for the frequency; (2) the sound pulse theory, by which each vortex on hitting the edge starts off another from the slit; this, incidentally, requires integral values of j; (3) the escapement theory, whereby any casual deviation of the jet from the forthright direction sets up a pressure increase in the surrounding air on one side of the wedge which, when it has grown to a sufficient magnitude, forces the jet back towards the other side, so making it pendulate. Brown himself favours the third (Wachsmuth's) theory with some modifications, but admits that more must be learnt of the fundamental properties of fluid jets before one can hope to evolve a quantitative theory of edge tones. The third investigation on jet tones, which bears an interesting relationship to the other two, forming as it does an intermediate case between the aeolian and the edge tone, is to be found in a paper by Carrière[51], who placed a wire about 2 mm. in diameter parallel to the slit from which an ignited gas jet was issuing. He found an approximate relation between the pitch of the tone produced, the velocity of efflux (or more precisely the square root of the gas pressure) and the distance between slit and wire, similar to that common to edge tones. The experiment seems to the writer to be suggestive in respect of the escapement theory which Brown endorses. If the wedge against which the pendulation is set up can be replaced by a thin wire without upsetting the mechanism of the edge tone, how can the postulated differences of pressure between either side be maintained when there is free communication between the two sides of the jet by way of the far side of the

wire? Carrière found that the wire itself vibrated if it happened to be in resonance with the vortex system. The matter is deserving of further research without the unnecessary complication of the flame, which introduces uncertain temperature fields.

Wind instruments form the next topic. Taber Jones[52] has treated the theory of the Haskell organ pipe, an open pipe into which a shorter closed pipe is inserted (figure 3) excited in the usual edge-tone method of flue pipes. In terms of the symbols shown on the figure, he proves that the wave-number k of the natural frequency is given by

$$q_2 \cot ka + (q_2 - q_1) \cot kb = q_1 \tan kl. \qquad \ldots\ldots(4)$$

The system of acoustic impedances is that of a stopped pipe with two open pipes in series branched upon the open end, and serves to produce bass notes in a comparatively restricted space, more so, in fact, than if a single stopped pipe were used. To make this formula agree with the experimentally observed frequencies, it is of course necessary to make certain assumptions with regard to end corrections. The end correction of an unflanged pipe forms the subject of a note by Bate[53] who shows that the interference between the two halves of a piston radiating into an

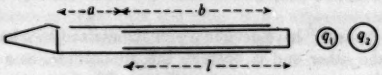


Figure 3.

infinite medium reduces the total conductivity in the ratio 2 : 1. It follows that the correction for a flanged tube of radius r is the same as that of an unflanged tube of radius $r\sqrt{2}$.

Lottermoser[54] has obtained the acoustic spectra of a number of organ pipes made of different materials to exhibit their influence upon the quality. He has thus demonstrated in physical detail a fact already known in part to organ builders, i.e. that the strongly damped heavy lead alloys overstress the lower partial tones, whereas with zinc and the more resonant metals amplification and even modulation of the higher partials may occur, producing a brighter and more strident tone which is exploited for trumpet and horn substitutes. Auger[55] has made measurements of the pitch of a reed pipe, showing how an increase in length of the resonating column lowers the pitch at first but is followed by a succession of jumps, so that a saw-tooth curve results when one plots frequency in terms of length. (This result is very similar to that which Trouton gave in a little known publication of some 30 years ago.)

The "howling tube" of Rijke, though scarcely a recognized musical instrument—even by modern composers—can conveniently be considered here, since Lehmann,[56] has made an extensive study of its mode of action. This apparatus consists of an open pipe held vertically and having a heated gauze fitted into its lower half. As long as the gauze is well above the temperature of the air which its convection current aspirates up the tube, it will give out a loud and—usually—strident sound.

The phenomenon was attributed by Rayleigh to a rectifying effect of the convection current on haphazard vibrations of the air column started by the draught rushing to enter the lower end of the tube—in the same way that one can produce sound from a hollow door key. When the *particle* velocity is directed upwards, cold air is brought on to the gauze, receives heat and is compressed, whereas a downward motion merely brings the already heated air back to the gauze. Thus the greatest heat transfer and maximum compression occur as the air in vibration moves up. When the gauze is in the lower half, this phase of the displacement occurs just before the maximum compression (because the air is moving towards the central node) and the motion is assisted; on the contrary, when the gauze is in the upper half. Lehmann has given mathematical expression to this theory and verified it in general save in this respect, that the output of sound does not increase indefinitely with the strength of the convection current, but reaches an optimum; with greater draughts the efficiency falls off. (It seems that this must be so, on the simple theory, for the rectifying effect will obviously diminish as the draught velocity approaches the maximum particle velocity in vibration.) The efficiency is measured in terms of the relation between the heat supplied to the gauze (electrically) and the pressure amplitude of the sound in the pipe. The latter was measured by a valve manometer (see below).

Andrade and Parker[57] have devised a pipe maintained by a loud speaker unit at one end, while the other end is open to the atmosphere, as a standard source suitable for making measurements of audibility. The intensity of the source is derived from observations of the amplitude of magnesium oxide smoke particles held in suspension inside the (glass) tube. The first author has shown that small smoke particles visible under a microscope acquire the full amplitude of the air in which they are mixed (within 1 per cent) if the frequency is below 1000 c./sec. Great care had to be taken to prevent convection perturbing the observations, and eventually the system was mounted in a thermostat. After considering possible methods of measuring particle size it was decided to observe their Brownian motion, i.e. the average displacement of a particle during a specific interval of time. The mean particle diameter of magnesium oxide particles produced by burning ribbon and fanning for 20 min. was found to be $3{\cdot}3 \times 10^{-5}$ cm., a size which is well able to follow the aerial motion without appreciable defect of amplitude. The frequency of the tone emitted by the diaphragm (working below its fundamental resonance) had to be accurately tuned to the second harmonic of the column of air, as the resonance peaks proved very sharp. When this was done the amplitude at the internal antinode over the working range (8 to 20×10^{-4} cm.) was found to be directly proportional to the current through the loud speaker. As a test of this source it was mounted on top of a building facing another across a court, the floor of which was padded with sawdust to prevent reflections. A listener stationed himself on the roof of the (somewhat lower) building opposite with his head in a specified position and his threshold of audibility was determined in terms of the energy radiated to the distant station from a "piston" source having the measured amplitude at the open end of the pipe, this being the same as that at the point one-third of the way along the pipe

from the diaphragm where the smoke particles were viewed. The averages of the minima of audibility of five auditors at two frequencies (410 and 646 c./sec.) agree with those of Fletcher and his collaborators. Two other measurements of amplitude in aerial vibrations are to be noticed. Schiller and Costelliz[58] have experimented with a couple of sound probes, one a "rectifying valve" consisting of a pinhole tube of the Barus type in which a steady pressure is built up, proportional to the pressure amplitude outside, and the other a hot-wire anemometer mounted behind a hole in a tube into which the sound waves penetrate. Lehmann[59] has constructed a valve manometer of the Kundt type and compared its indications with that of a Rayleigh disk in a tube. Such an apparatus has a considerable lag and is only suited to low frequency measurements. The valve is held on its seat by a light spring which opens in the half period during which the pressure is in excess but closes in the other half.

Timotika[60] has worked out the interference which the motion of a Rayleigh disk should experience in virtue of the proximity of the walls of the resonance tube in which it is placed; a calculation which recalls the corresponding interference of the channel walls on the forces of an aerodynamic model in a wind tunnel. The correction to the moment on the disk is expressed in a series of powers of d/D, the ratio of the respective diameters of the disk and tube. Some experimental confirmation of the theory is forthcoming, which shows that the wall effect first becomes appreciable when $d/D = 0{\cdot}25$ and then rises rapidly. At $d/D = 0{\cdot}5$, however, the moment is only about 12 per cent larger than in an infinite medium.

Pohlman[61] has treated the directional effect of a sound field on a suspension of aluminium flakes in xylol. In such fields these particles behave like miniature Rayleigh disks and begin to set themselves perpendicular to it, as well as migrating towards the antinodes. In consequence of this lining-up there is a progressive change in the amount of light which gets through the suspension viewed along a normal to the wave-front, similar to that which may be got by applying an alternating *electrostatic* field to a suspension of charged flakes, e.g. graphite oxide in oil.

§ 4. ACOUSTIC IMPEDANCE

Apart from a paper by Lindsay[62] which gives a mathematical treatment difficult to summarize of the propagation of sounds along an acoustic filter containing a finite number of impedances in place of the infinite line which, for the sake of simplicity, is usually assumed in such calculations, the only work to be noticed under this heading is the description of an acoustic impedance bridge, constructed and tested by Robinson[63]. To take the place of the ratio arms of a Wheatstone Bridge, reactances were introduced consisting of wide tubes of length l and area S for which the impedance was taken to be $-i\rho c/S.\cot 2\pi l/\lambda$ (ρ being density of air, c velocity and λ wavelength of sound used). The source of sound was a loud speaker (outside the room) connected to one corner of the diamond network, the opposite corner being a dead-end or earthed into the atmosphere. The arrangement is shown in figure 4. The loud speaker feeds to a junction box J, then follow the ratio arms and

beyond the connecting pieces P, P, the unknown impedance (shown as a horn) and the known impedance (a tube of adjustable length). Rubber tubes T, T lead to the galvanometer—a microphone M, with compensating tubes and pistons C, C permitting an adjustment of the acoustic impedance on either side of the diaphragm. The apparatus was first used to measure standard forms of impedances such as the Helmholtz resonator. Difficulty in deciding what end correction to ascribe to an orifice which does not debouch into the open air, but into a coaxal pipe of different section, led eventually to the use of the bridge to determine such corrections. Even when the orifices were used alone, without the reservoir, higher values for the conductivity were obtained than the theoretical value, which for a circular hole in a

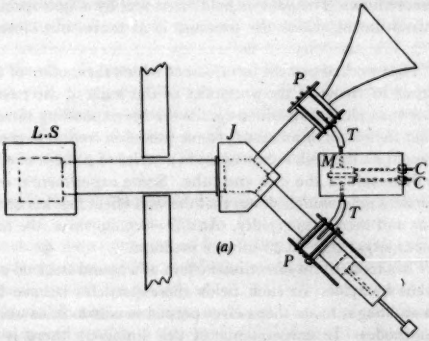


Figure 4. (Reproduced with the kind permission of the Editor of the *Philosophical Magazine*.)

thin plate is equal to the diameter. It was found that the conductivity of a hole of diameter d in a tube of diameter D could be represented by

$$C=\frac{0{\cdot}787\,d}{(1-d/D)^{1{\cdot}895}}$$

if it formed a constriction in a tube and $C=d\,(1+d/D)^{1{\cdot}19}$ if it were a terminal.

Finally a simplified form of the bridge was built up consisting of a single tube 10 in. long with a single telephone source on a branch at the centre. Two pressure tubes were brought to the axis of the main tube, one either side of the T junction, 1 in. from the ends of the main tube. The differential microphone was then attached to the two pressure tubes and unknown and known impedances joined to the ends of the main tube. The bridge has proved to be convenient and accurate for determining acoustical reactances, and work on resistances of the slot type is now in progress.

Morrical[64] describes an interesting modification of the tube method of measuring the acoustic properties of materials. It is usual in such apparatus to produce standing waves along the tubes between the loud-speaker source and the specimen and to

explore the pressure or displacement pattern along it. Morrical, however, prefers to set the specimen at 45° to the axis of the main tube and cast the reflected radiation into a branch tube at 90° to the first, picking up the reflected energy on a condenser microphone. In order to prevent further reflections from the ends of the main and the branch tubes, they are widened out in the form of exponential horns, so that the impedance of the bell-shaped end is matched to that of the open air and negligible rebound ensues. The microphone was placed at the open end of the branch horn. Three readings were taken: (1) with the specimen in place, giving a pressure due to the reflected + the diffracted energy from the main tube; (2) with a thick glass plate at the junction giving a reading equal to the incident energy + the diffracted energy; (3) with no reflector giving diffraction alone. It was assumed that the latter remained constant. Then it was only necessary to subtract (3) from (1) and (2) and take the ratio of the remainders to get the reflection coefficient. Actually the pressure in the main tube does not remain constant, as we have tacitly assumed, but varies with the load on the loud-speaker. Allowance for this was made by taking readings with a miniature condenser microphone pushed into the loud speaker throat.

§5. SPEECH AND HEARING

Vowel sounds continue to interest both physicists and physiologists. Gemelli[65]—who gives his paper an abstract in Latin in the antique mode—describes further work on the characteristic structure of the chief (Italian) vowels *a*, *i*, *u*, directing notice in particular to the extent of variation from period to period in the wave form of such speech sounds. His conclusions may be shortly summarized as follows:

a is characterized by important high frequency components so related that they interfere and produce a *smorzando* effect between each impulse of breath from the vocal cords.

i, on the other hand, finds most of the energy in the fundamental and first overtone, but has also some weak high frequency components, independent of the pitch of intonation (fundamental).

u contains mainly low frequency components, some of which beat together.

Vierling and Sennheiser[66] have also recorded vowel wave forms with a view to detecting the essential differences between short and long vowels. The principal conclusion—already commonplace in the phonetic world—seems to be that no question of relative duration arises, only one of type. Farnworth[67] tackles the subject rather from the psychological aspect—a very important one. He presents a series of pure tones to college subjects and a list of a dozen typical vowels to one of which they are to assimilate the tone heard. It is rather difficult to draw satisfying conclusions from this type of test with regard to the resonant characteristics of vowels though the author succeeded in arranging them in order of natural pitch.

The brothers Trendelenburg[68] have the undoubted advantage of forming a family team—physicist and physiologist—admirably adapted to this kind of research. In their latest work they find (by octave filters) that the opening and shutting of the vocal cords (which occurs of course once within the fundamental period of the

vowel) is accompanied each by a peculiar group in the region of 7000 c./sec. which accounts in part for the recorded wave form. Another paper exhibits simultaneous records of the vibrations of the larynx and breast walls and the sound produced (in the external air) when a vowel is explosively enunciated (*coup de glotte*) in the manner approved by some singers. No appreciable lag between the two records could be detected, contrary to the assertion of certain voice specialists. For the understanding of the human vocal mechanism and for phonetics generally it is desirable to make simultaneous records of the form of the vocal cavities when they are held in proper position for producing speech sounds, and of the sounds themselves when voiced. Curry[69] describes an electrical timing device by which an X-ray exposure of the throat may be made at any time interval after the initiation of speech, thus avoiding the uncomfortable holding of mouth positions necessary with the older techniques, in which one could never be certain that the X-ray snapshot corresponded to any part of the recorded sound. It is possible to arrange that the relay operating the X-ray exposure is worked by an excess of current in the microphone circuit, so letting the initiation of speech or of some specially emphasized speech sound take the radiograph automatically.

Another useful piece of apparatus is described by Obata and Kobayashi[70]; it is an electrical analyser using thyatron tubes in which each discharge produces a pulse whose intensity is proportional to the frequency of the input current. The input is taken from a gramophone pick-up, while the variations in output are recorded on a moving film to a logarithmic scale of pitch. The apparatus is thus able to follow rapid changes of pitch during an instrumental solo or of intonation during a speech.

On *hearing* there are, as usual, quite a number of interesting papers. In the last *Report* (p. 63) we described the experiments of Hallpike, Hartridge and Rawdon-Smith[71] to test whether the sudden reversal of phase (by 180°) in the sinusoidal output of a siren caused an interruption in the response in the ear. They have now improved the technique by which the phase change is produced and are able to substantiate the intervention of beats in the response of the auditory tract of a decerebrate cat. On the other hand, the electric stimulus in the cochlea (which Weaver and Bray had previously picked up) was found to follow the objective sound wave form with remarkable fidelity. They conclude that the Weaver-Bray phenomenon is unconnected with any resonant response in the cochlea.

Stevens, Volkmann and Newman[72] have made experiments to establish a subjective scale of pitch. Since a considerable effect of loudness on perceptual pitch has been proved in recent years, it is no longer sufficient to measure pitch in terms of frequency alone. They proceeded on similar lines to those through which scales of loudness have been built up. The auditor was required to adjust the frequency of a tone until it seemed to have half the pitch of a standard tone, presented at a constant loudness of 60 db. Some observers at once objected that pitch was not something they could halve, but soon were able to say when the comparison "felt like half", though some acquired this faculty nore readily than others, and trained musicians with their pre-conceived notions of the octave were naturally not good

subjects from the point of view of those directing the experiments. The results are shown on figure 5, the dotted line being that which would be obtained if "half the pitch" always meant the sub-octave. On the same figure the chain line represents the perceived size of a sub-octave at different frequencies. A musical interval in fact becomes (subjectively) larger as it is carried up to the fourth octave above middle C. Steinberg[73] has collected the available data on the resonant positions along the basilar membrane which should correspond to various levels on the audible frequency scale; and these have been used by Fletcher and Munson[74] to correlate loudness with intensity of stimulus per unit length of the membrane. Steinberg and Gardner[75] have measured the loudness sensitivity of a number of cases of impaired hearing and find that, above their threshold of audibility, many such persons possess normal sensitivity. This explains why many partially deaf people complain if they are shouted at.

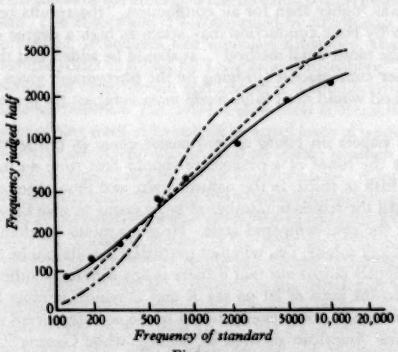


Figure 5.

A number of papers are devoted to the non-linear characteristics of the ear. Stuhlman[76] has constructed a model of the middle ear and shown that distortion is of necessity produced by the ossicles in passing a simple harmonic motion of the stapes to the inner ear. Newman, Stevens and Davis[77] have used the potential set up in the cochlea by sound applied to the outer ear to detect the incidence of subjective harmonics in the response to pure tone. An electrode applied to the ear of a cat or guinea pig was connected to a wave-analyser. The harmonics appear at that value of the intensity (40 to 60 db.) at which the fundamental response ceases to be proportional to the stimulus. They have also detected combination tones in the cochlear response when the simple harmonic motions were applied. Trimmer and Firestone[78] have also been testing the ear in relation to combination tones, and claim to have found an effect dependent on the relative phase of the constituents; another result of the non-linearity of the response of the ear. Stevens[79], considering possible theories to account for the fact that a sound may be heard when an alter-

nating electrical current is applied to the ear, favours that one which holds the effect to be the converse of the Weaver-Bray effect noted above, i.e., that mechanical vibrations are set up the ear by the a.c. Watson[80] has carried out a careful research on the hearing of speech by bone conduction, in which a special vibrator was constructed for stimulating the frontal bone of the skull through a duralumin rod which pressed upon the forehead and was actuated by a small induction coil to which amplified currents picked up by a microphone were led. The observer and the vibrator were placed in a sound-proof room and articulation tests carried out (using gramophone records as source of speech) at various sound levels and with different contact buttons on the piston. Alteration in the latter caused changes in the mean pressure of the rod upon the head. The condition of the head affected the results. Ultimately it was decided "mouth closed, teeth in contact" gave the optimum conditions of intelligibility and comfort to the patient. Although the mean threshold was higher than for air conduction, "the results revealed that the hearing of speech by bone conduction may attain as high a degree of perfection as that attained by the more usual method". It should be added that the experimental vibrator was rather cumbersome (judging by the photograph given, about half the size of the head) and would need to be made more compact for practical application and portability.

A number of papers on *tuning of instruments* come in this section. Harker[81] recommends that the tuner of a pianoforte should first tune an interval between two notes, e.g. a fifth or third, to the natural scale, and then raise or lower the frequency of one until the requisite number of beats ensue to give the correct relation of frequencies on the equi-tempered scale. He gives tables for doing this. On the vexed question of key colour, i.e., whether particular moods can be associated with certain keys, he rightly points out that if there is any difference other than absolute frequency between the keys of let us say A and C major, it must be because the instrument in question is not properly tuned to the equi-tempered scale. White[82] animadverts on the American method of tuning, while Greene[83] has advanced scientific evidence to combat another popular fallacy. It has often been held by violinists and singers that, when giving solo performances unfettered by dependence on keyed instruments, they adhere to the natural and not the equi-tempered scale. Greene has tested a number of violin virtuosi and the answer is that they conformed to neither scale but evolved an individual compromise between them.

Bachem[84] has made an interesting study of absolute pitch. Genuine possessors of this faculty could at once name a note presented to them on any instrument, apart from occasionally putting it in the wrong octave, whereas others, even when musically trained, had to think for some seconds, during which time they made mental comparisons with their preconceived notion of A or C, as the case may be, and were 5 to 9 semi-tones out in the final guess. Other possessors of a "quasi-absolute" pitch judged it from their vocal sensations in humming the note. The fact that only 7 per cent of the subjects possessed absolute pitch in respect of all instruments, while 44 per cent possessed it in respect of their chosen instrument only, suggests that it is primarily a feat of memory which is involved.

§ 6. TECHNOLOGY

Sound insulation. Quite a number of papers this year deal with the acoustic properties of walls and partitions. Constable and Aston[85] report sound insulation measurements on single and multiple partitions made at the National Physical Laboratory to show in particular how the reduction in transmission varies with the weight per ft.^{2} at a number of frequencies between 200 and 2000 c./sec., while Davis and Knowler[86] point out that when a layer of air is sandwiched between solid partitions for the purpose of insulation, the coupling between the slabs afforded by the air sac may limit the efficacy of the device. Lauffer[87] has also treated the sound absorption produced by a diaphragm separated by a layer of air from a massive wall. The frequency of maximum attenuation is in the bass and depends on the mass of the plate and the thickness of the air film, while its value is determined by the amount of damping therein. The absorption is accompanied by drumming of the plate. If, however, it is perforated, it behaves like a porous substance, absorbing the high frequencies. A similar result is found by Lubcke and Eisenberg[88]. Janovsky and Spandock[89] describe the construction and testing of a sound-damped chamber at the Siemens-Halske laboratories. The outer shell is built of Heraklith bricks; inside this is a 10 cm. layer of slag wool, followed by an equal layer of wadding on to which further pads of wadding are placed end-on with air spaces between them. Grids on the floor allow of the passage to and fro of the experimenters. A loud speaker occupies a space in front of the door. Tests of the intensity distribution of sound down the longitudinal axis of the room show that at first the sound spread in truly spherical form with inverse square reduction of amplitude. Near the walls, however, although no reflections were detected, the fall of amplitude seemed to follow at a greater rate than theory would indicate. This they attribute to the diffraction which the sound waves suffer in passing over the surface of the absorbing walls. The velocity in such locations is in fact less than in the open air, so that a sort of acoustic mirage takes place in the vicinity of the wall. The depth of the diffraction zone is proportional to the wavelength of the sound and so is less extensive at high frequencies. The total absorbing power of the room is 98 per cent for 1000 c./sec. and 99·4 per cent at 4000 c./sec. So the experiments show that to construct a room to simulate free field conditions one must have regard not only to minimize reflection but also diffraction effects caused by the walls. A paper by Sivian[90] deals with an analogous problem, viz. the attenuation of sound by its passage through a conduit lined with absorbent material. The problem is treated for the conditions of both rigid and mobile lagging in terms of an acoustic line with uniformly distributed series impedance Z_1 per unit length, together with shunt admittance $1/Z_2$ per unit length in branches; such that

$$Z_1 = \frac{i\omega\rho}{A}; \quad \frac{1}{Z_2} = \frac{i\omega A}{\gamma p} + \frac{P}{z}.$$

Here A is the area and P the perimeter of the duct; z is the motional impedance per unit area of the lining and the other symbols have the usual significance in acoustics. Three cases are worked out: (1) in which the wall impedance is a pure resistance,

(2) in which it has also stiffness, (3) in which it has also inertance. In the third case the structure acts as a high-pass filter. Finally the effect of allowing the lining to indulge in transverse vibration, propagated along the conduit, is considered. The propagation constant is then a function of that for waves in the isolated lining and of that for the pure aerial waves, and the coupling between the solid and fluid vibrations.

Meyer and Keidel[91] describe an apparatus by means of which the sound insulation of insulating supports for machinery can be measured. The vibration is transmitted through the specimen from an iron mass wound with an induction coil, under a freely hanging electro-magnet carrying the alternating current which initiates the vibration. The impulse transmitted through the test piece is picked up by a quartz receiver in direct contact with it. The potential excited piezo-electrically in the quartz is then measured on a valve voltmeter. Response curves from frequencies 50 to 200 c./sec. are then quickly obtained, for which the insulating properties and likely resonances in the material are quickly read off. A model mechanical filter consisting of a number of metal disks (each weighing 600 g.) separated by 10-mm.-thick rubber pads was found by this method to have a sharp cut off at 400 c./sec. and two resonances at 60 and 180 c./sec. (The electrical analogue of this low-pass filter is a series of inductances in line, with condensers in the branches.)

Singing propellers. Since the introduction of propellers made of bronze in ships a few years ago, much interest has been shown by engineers in the obnoxious noise which such propellers may develop when running. This interest culminated in a very detailed account given by Hunter[92] before the North-East Coast Institution of Engineers of cases with which his firm has been concerned. The trials—in more than one sense of the word—recounted in his paper form a very valuable mass of evidence on which a discussion of this phenomenon may be based. An uncanny feature of these cases is that the trouble often does not develop until the ship goes to sea, the basin trials with the propeller submerged to different depths not leading the owners to suspect anything wrong, while of propellers made to identical patterns and as far as instruments can detect of equal size, one may become a singer and the other remain dumb.

The acoustic features of the phenomenon may be summarized as follows:

(1) The sound is rather difficult to describe but is somewhat like that which would be expected if the blade received a blow with a moderately hard hammer once in each revolution. The low pitched components of the grouching noise thus resulting are rapidly damped but the high pitched components remain as a metallic ringing from one impulse to the next.

(2) The noise is only blatant over a certain range of engine speeds, reaching a maximum and then waning as the r.p.m. are increased through the noisy range.

(3) The pitch of the continuous ringing may be identified with one or other of the torsional vibrations of the blade. The pitch of the most favoured partial tone changes with the r.p.m.

(4) The noise is only important when an alloy of suitable elastic properties, such as bronze or stainless steel, is used. Secular changes in the structure of the alloy may abate or aggravate it.

(5) The same effect may be produced by a change of helm or of physical properties of the water.

In the minds of many naval engineers the singing of a propeller is closely bound up with the corrosion which it often experiences in use. To the extent that corrosion changes the physical properties and configuration of a propeller, it is bound to have an influence on its acoustic properties, but it is not yet proven that the noise is due to the same cause as the corrosion, i.e. the liberation of dissolved gas as the blades lash the water. There is also the difficulty of explaining fact (3) on these lines, since one cannot conceive an obvious reason why the frequency of degassing should be a function of the speed. A more likely cause is to be found in the aeolian vortices formed behind the blades and which trace out helical paths in the wake as the blades rotate. The vibrations so forced on the blades would be considerably enhanced if they happened to coincide with one of the natural frequencies of the blade in a transverse or flexural mode, and the collapsing of the liberated gas into these vortices might further add to the energy available for exciting noise. In this connexion some experiments of Stowell and Deming, illustrated in the last *Reports*, are pertinent. They idealized a propeller into a uniform rod of circular section, pierced at its mid-point by an axle about which it was revolved by an electric motor, while they picked up the sound produced on a microphone. The sound spectra so obtained (cf. volume **4**, figure 1, p. 58) illustrate clearly how the pitch of the hum shifts from one set of partial tones to another as the r.p.m. of the rod is increased. The highest component in each case corresponds to the frequency of production of vortices at the fastest moving portion of the rod, i.e. the tips. This would account for the possibility of exciting vibrations at quite a range of speeds. At the same time, disturbances from upstream of the propeller, e.g., those produced at the stern post which preserve for a time their relative positions, produce additional impulses on the screw co-periodic with the revolutions. Such probably give rise to the highly damped vibrations at low revolutions. In the view of the writer[93], the phenomenon of the singing propeller comes within the scope of relaxation oscillations. These are produced whenever a system receives regular impulses sufficiently detached to allow it to oscillate in one or other of its degrees of freedom with diminishing amplitude between each impetus. This would account for (1). The permissible vibrations of the blades of a scale model flat-bladed propeller have been studied by Hunter in the familiar method of Chladni figures, bowing at various places to exhibit the modes of vibration when the real propeller is excited at corresponding points. These model experiments were done in air and the working stresses were absent, so that as the author points out, the sound was far more musical and tolerable than that which occurs at sea; under no conditions was the distressing grinding or grating of noisy propellers set up in the model.

Sound direction finding. Although there are no papers which require detailed treatment under this heading, it is perhaps apposite, in view of widespread interest in air raid precautions, to give some account of the methods of technical listening by which the positions of hostile aircraft are ascertained. There are two methods commonly employed, but the one which is favoured in this country possesses

the advantage of its greater mobility, if not portability. This depends on the human faculty of binaural hearing, by which we are able to tell from which direction sound is coming towards us and in particular when it lies directly in front (or directly behind, since one cannot discriminate between fore and aft with this sense). The possibility of using this faculty, with instrumental aids, for the spotting of the emplacement of guns had been considered by the military towards the end of last century, although it was not until the aeroplane became a potential menace that full use was made of it. Indeed the late Lord Rayleigh had in 1876 pondered the means by which binaural listening could inform a person of the direction of arrival of sounds at his head, and had separated out the two pertinent effects, viz. the intensity difference at the two ears and the phase difference. He decided that in the lower frequency range the phase difference must be the discriminating factor. Indeed when one reflects that the head can only form a considerable shadow over the further ear to sounds which have the half wavelength equal to the diameter of the head, one concludes that it will only be above 800 c./sec. that the intensity difference can make a notable contribution to the detection of direction. Actually, though the sounds made by an aeroplane have a multiple origin, propeller noises, aeolian tones of struts, exhaust puffs, etc., the important region for listening purposes lies between 300 and 600 c./sec. When an electrical circuit intervenes between the artificial ears and the human ones, the rest of the noise spectrum is generally removed in a band-pass filter. The ways in which the detecting apparatus improves on the unaided ears are twofold: (1) trumpets or mirrors collect the sound over a wider area than the human ear and concentrate it on to the throat or focus of the apparatus as the case may be, (2) the collectors are spaced at a greater distance apart (usually about 12 ft.). The object of the former is obvious. The latter increases the sensitivity of direction fixing. The auditory resolving power is a function of the least time difference between arrival of compressions which the ears can detect. The actual sensitivity can then be increased in proportion to (*a*) the wavelength of the sound, (*b*) the separation of the artificial ears. In practice there are four of these arranged in ⊣ fashion. The whole apparatus is pivoted on a universal joint at the junction of the arms of the tee. The two collectors at the ends of the horizontal arm serve to determine the bearing on a horizontal projection, and are under the control of one observer. The other finds the elevation, having the trumpets at the end of the upright connected severally to his ears. When they have finished spotting, the plane of the framework should lie at right angles to a line joining the apparatus to the sound source, in the absence of refraction *en route*. The observers naturally need training in the use of the binaural faculty, and in working the apparatus without interfering with each other's movements, but with practice, it is said to be possible to detect within 0·1 degree with the British pattern apparatus. At night the movement of a searchlight beam is often co-ordinated with the direction finder. Of course it is important that unwanted noises shall not penetrate the collectors. The sounds set up when the apparatus is exposed to the wind are troublesome, especially aeolian tones of supports, and edge tones formed in the trumpets which may excite the air cavities into resonant vibration (cf. the sounds heard in holding a shell to the ear); but these can be

minimized by padding the sharp edges of the trumpet frames so that the adventitious noises do not mask the distant source which it is required to detect. The second type of direction finder does not use the binaural faculty. All that it asks of the ear is to judge the loudest of a set of sounds presented in turn to it, and this operation can be performed instrumentally, if it is desired to make the whole operation objective. A single large reflector several metres across collects the sound and delivers it to a set of some 30 electrodynamic microphone units disposed about the focal plane. These must be nicely matched in characteristics. Most of the sound energy in a parallel beam incident upon the mirror will then be concentrated in one or two of the pick-ups. Actually the focal plane in the type of mirror employed is a sphere. In the compensator which measures the intensity of the current from each microphone, contacts are similarly disposed about a smaller sphere over which the operator passes a switch to pick out the maximum response, the azimuth of the corresponding contact can then be seen on the sphere. The apparatus, though less mobile than the first type, has this advantage that the recording instruments can be at a distance from the mirror and in a more protected site. For submarine direction-finding under the sea, it is the only serviceable type, the binaural detector being too subject to casual sounds. In the naval type the microphones are disposed only in a horizontal plane, on half-ellipses, one each side of the vessel which is picking up the noise. The bearing of the source of sound can then be read off directly on a dial in the instrument room.

Bug-hunting by sound. An interesting application of microphone technique sponsored by the Forest Products Research Laboratory to the location of destructive larvae in wood is described by Colebrook[94]. Unless the experiment is done in a sound-proof chamber, the sound-level of the environment is higher than that due to the insects. Even so the gain required in the amplifier is large enough to make background noise obstreperous. In the case discussed by Colebrook, sounds between 1000 and 5000 c./sec. were sought for in a piece of wood suspected of harbouring the death-watch beetle. An intermittent rattle, which could be brought to loud-speaker intensity, was ascribed to the larvae. Incidentally, the author mentions that Schwarz, Kranz and Sicke used a similar method to detect the larvae of the house longhorn two years ago. It is hoped that it will be possible eventually to apply the method to the suspected timber of buildings, without the need to test piecemeal in a sound-proof box.

REFERENCES

(1) Hodge, A. H. *J. Chem. Phys.* **5**, 974 (1937).
(2) Hubbard, J. C. and Hodge, A. H. *J. Chem. Phys.* **5**, 978 (1937).
(3) Eucken, A. and Nümann, E. *Z. phys. Chem.* B, **36**, 163 (1937).
(4) Warner, G. W. *J. Acoust. Soc. Amer.* **9**, 30 (1937).
(5) Itterbeek, A. van and Mariëns, P. *Physica*, **4**, 207 and 609 (1937).
(6) Metter, I. M. *Phys. Z. Sowjet.* **12**, 233 (1937).
(7) Bazulin, P. *C.R. Acad. Sci., U.R.S.S.*, **14**, 273 (1937).
(8) Korolev, F. A. *C.R. Acad. Sci., U.R.S.S.*, **15**, 35 (1937).
(9) Bergmann, L. and Goehlich, H. J. *Phys. Z.* **38**, 9 (1937).
(10) Nomoto, O. *Proc. Phys. Math. Soc., Japan*, **19**, 337 (1937).

(11) Claeys, J., Errera, J. and Sack, H. *Trans. Faraday Soc.* **33**, 136 (1937).
(12) Schaafs, W. *Z. Phys.* **105**, 658 (1937).
(13) Wyss, R. *Helv. phys. Acta*, **10**, 237 (1937).
(14) Bär, R. *Helv. phys. Acta*, **10**, 332 (1937).
(15) Otpushtshennikov, N. F. *Phys. Z. Sowjet.* **12**, 736 (1937).
(16) Fox, F. E. *Phys. Rev.* **52**, 973 (1937).
(17) Rao, B. V. *Nature, Lond.*, **139**, 885 (1937).
(18) Lucas, R. *J. Phys., Radium*, **8**, 41 (1937).
(19) Brandt, O. *Kolloidzschr.* **81**, 2 (1937).
(20) Brandt, O., Freund, H. and Hiedemann, E. *Z. Phys.* **104**, 511 (1937).
(21) Landau, L. and Rumer, G. *Phys. Z. Sowjet.* **11**, 18 (1937).
(22) Rzevkin, S. N. *C.R. Acad. Sci., U.R.S.S.*, **16**, 267 (1937).
(23) Hiedemann, E. and Hoesch, K. H. *Z. Phys.* **104**, 197 (1937).
(24) Schreuer, E. and Osterhammel, K. *Z. Phys.* **107**, 44 and 560 (1937).
(25) Hiedemann, E. and Osterhammel, K. *Z. Phys.* **107**, 273 (1937).
(26) Hiedemann, E. and Schreuer, E. *Z. Phys.* **107**, 463 (1937).
(27) Bär, R. *Helv. phys. Acta*, **10**, 311 (1937).
(28) Baumgardt, E. *C.R. Acad. Sci., Paris*, **204**, 416 and 751 (1937).
(29) Hiltscher, R. *Z. Phys.* **104**, 672 (1937).
(30) Ostrovsky, E. P. *C.R. Acad. Sci., U.R.S.S.*, **14**, 491 (1937).
(31) Brohult, S. *Nature, Lond.*, **140**, 805 (1937).
(32) Tyler, E. *Phil. Mag.* **24**, 665 and 905 (1937).
(33) Salceanu, C. *C.R. Acad. Sci., Paris*, **205**, 1219 (1937).
(34) Trimmer, J. D. *J. Acoust. Soc. Amer.* **9**, 162 (1937).
(35) Dahl, H. and Denk, O. *Nature, Lond.*, **139**, 550 (1937).
(36) Wood, A. B., Tomlinson, G. A. and Essen, L. *Proc. Roy. Soc.* A, **158**, 606 (1937).
(37) Backhaus, H. *Z. tech. Phys.* **18**, 98 (1937).
(38) Meinel, H. *Akust. Z.* **2**, 22 and 63 (1937).
(39) Minnaert, M. and Vlam, C. C. *Physica*, **4**, 361 (1937).
(40) Saunders, F. A. *J. Acoust. Soc. Amer.* **9**, 81 (1937).
(41) Vierling, O. *Z. tech. Phys.* **18**, 103 (1937).
(42) Kock, W. E. *J. Acoust. Soc. Amer.* **8**, 227 (1937).
(43) Lehmann, K. O. *Z. tech. Phys.* **18**, 304 (1937).
(44) Pavlik, B. *Z. Phys.* **107**, 458 (1937).
(45) Pavlik, B. *Ann. Phys., Lpz.*, **28**, 353 and 632 (1937).
(46) Waller, Mary. *Proc. Phys. Soc.* **49**, 522 (1937).
(47) Hayes, H. C. *J. Acoust. Soc. Amer.* **8**, 220 (1937).
(48) Jones, A. T. *J. Acoust. Soc. Amer.* **8**, 199 (1937).
(49) Lehnert, R. *Phys. Z.* **38**, 476 (1937).
(50) Brown, G. B. *Proc. Phys. Soc.* **47**, 703 (1937).
(51) Carrière, Z. *Rev. Acoust.* **5**, 138 (1936).
(52) Jones, A. T. *J. Acoust. Soc. Amer.* **8**, 196 (1937).
(53) Bate, A. E. *Phil. Mag.* **24**, 453 (1937).
(54) Lottermoser, W. *Akust. Z.* **2**, 129 (1937).
(55) Auger, W. L. *C.R. Acad. Sci., Paris*, **204**, 1169 (1937).
(56) Lehmann, K. O. *Ann. Phys., Lpz.*, **29**, 527 (1937).
(57) Andrade, E. N. da C. and Parker, R. C. *Proc. Roy. Soc.* A, **159**, 507 (1937).
(58) Schiller, P. E. and Costelliz, H. *Akust. Z.* **2**, 11 (1937).
(59) Lehmann, K. O. *Z. tech. Phys.* **18**, 309 (1937).
(60) Timotika, S. *Proc. Phys. Math. Soc. Japan*, **19**, 329 (1937).
(61) Pohlman, R. *Z. Phys.* **107**, 497 (1937).
(62) Lindsay, R. B. *J. Acoust. Soc. Amer.* **8**, 211 (1937).
(63) Robinson, N. W. *Phil. Mag.* **23**, 665 (1937).
(64) Morrical, K. C. *J. Acoust. Soc. Amer.* **8**, 162 (1937).
(65) Gemelli, A. *Comm. Pont. Acad. Sci.* **1**, 1 (1937).
(66) Vierling, O. and Sennheiser, F. *Akust. Z.* **2**, 93 (1937).

(67) FARNWORTH, P. R. *J. Acoust. Soc. Amer.* **9**, 152 (1937).
(68) TRENDELENBURG, F. and W. *S.B. preuss. Akad. Wiss.* **13**, 127; and **20**, 265 (1937).
(69) CURRY, R. *Rev. Sci. Instrum.* **8**, 382 (1937).
(70) OBATA, J. and KOBAYASHI, R. *J. Acoust. Soc. Amer.* **9**, 156 (1937).
(71) HALLPIKE, C. S., HARTRIDGE, H. and RAWDON-SMITH, A. F. *Proc. Phys. Soc.* **49**, 190 (1937).
(72) STEVENS, S. S., VOLKMANN, J. and NEWMAN, E. B. *J. Acoust. Soc. Amer.* **8**, 185 (1937).
(73) STEINBERG, J. C. *J. Acoust. Soc. Amer.* **8**, 176 (1937).
(74) FLETCHER, H. and MUNSON, W. A. *J. Acoust. Soc. Amer.* **9**, 1 (1937).
(75) STEINBERG, J. C. and GARDNER, M. B. *J. Acoust. Soc. Amer.* **9**, 11 (1937).
(76) STUHLMAN, O. *J. Acoust. Soc. Amer.* **9**, 119 (1937).
(77) NEWMAN, E. B., STEVENS, S. S. and DAVIS, H. *J. Acoust. Soc. Amer.* **9**, 107 (1937).
(78) TRIMMER, J. D. and FIRESTONE, F. A. *J. Acoust. Soc. Amer.* **9**, 24 (1937).
(79) STEVENS, S. S. *J. Acoust. Soc. Amer.* **8**, 191 (1937).
(80) WATSON, N. A. *J. Acoust. Soc. Amer.* **9**, 99 (1937).
(81) HARKER, G. F. H. *J. Acoust. Soc. Amer.* **8**, 243 (1937).
(82) WHITE, W. B. *J. Acoust. Soc. Amer.* **9**, 47 (1937).
(83) GREENE, P. C. *J. Acoust. Soc. Amer.* **9**, 43 (1937).
(84) BACHEM, A. *J. Acoust. Soc. Amer.* **9**, 146 (1937).
(85) CONSTABLE, J. E. R. and ASTON, G. H. *Phil. Mag.* **23**, 161 (1937).
(86) DAVIS, A. H. and KNOWLER, A. E. *Phil. Mag.* **23**, 154 (1937).
(87) LAUFFER, H. *Hochfrequenztechnik*, **49**, 9 (1937).
(88) LUBCKE, E. and EISENBERG, A. *Z. tech. Phys.* **19**, 135 (1937).
(89) JANOVSKY, W. and SPANDOCK, F. *Akust. Z.* **2**, 322 (1937).
(90) SIVIAN, L. J. *J. Acoust. Soc. Amer.* **9**, 135 (1937).
(91) MEYER, E. and KEIDEL, L. *Z. tech. Phys.* **18**, 299 (1937).
(92) HUNTER, H. *Trans. N.-E. Cst. Instn. Engrs Shipb.* **53**, 189 (1937).
(93) RICHARDSON, E. G. *Engineering*, **144**, 170 (1937).
(94) COLEBROOK, F. M. *J. Sci. Instrum.* **14**, 119 (1937).

INSTRUMENTAL AIDS FOR DEFECTIVE HEARING

By PHYLLIS M. TOOKEY KERRIDGE, M.Sc., Ph.D., M.R.C.P.
University College, London

§ 1. INTRODUCTION

Persons with normal hearing may use aids to hearing under special circumstances, as when listening to public speakers in large halls, or for localizing aircraft. But such cases will not be dealt with in the present article as information concerning them can be obtained from non-medical acoustic literature, or by experiments on normal human subjects. The problem of magnifying sound suitably for persons with defective hearing will be alone considered. This is a topic on which such meagre information as exists is widely scattered; but it is one of considerable technical interest and difficulty, and furthermore of great practical human importance. In the development of the subject medical men dealing with deaf people have been hampered through lack of knowledge of acoustics, and of facilities for carrying out a difficult research necessitating an elaborate technique which is not used in other branches of medicine. Physicists and instrument makers interested in the subject have been handicapped on their part by insufficient information on the physiology and pathology of the ear. In this paper more particulars of the ear have been given than can be obtained readily from physical textbooks, and an attempt has been made to avoid medical terminology as far as possible. A classified bibliography may help readers wishing to study the subject in more detail.

§ 2. THE EAR AND THE AUDITORY NERVOUS PATHWAY IN HEALTH AND DISEASE

The effect of sound on the ear. The structure of the ear and the route of the electrical impulses from the ear to the brain are illustrated very diagrammatically in figure 1.

Sound vibrations in the external ear cause movements of the ear drum, which in turn cause vibration of the tiny bones in the middle ear, the oval window of the inner ear, the perilymph fluid of the cochlea, the endolymph fluid between Reissner's membrane and the basilar membrane, and the Organ of Corti on the latter. To the bases of the hair cells of this Organ are attached nerve fibres which communicate with nerve cells in the central part of the cochlea. A second series of branches from these spiral ganglion nerve cells carry the impulses on to the brain.

These nerve branches join some from the neighbouring semicircular canals, and they emerge together as a thin cord within the bony skull and enter the brain. The auditory fibres soon end at nerve junctions within the brain. The electrical changes of potential or impulses they transmit are carried on by other nerve fibres which

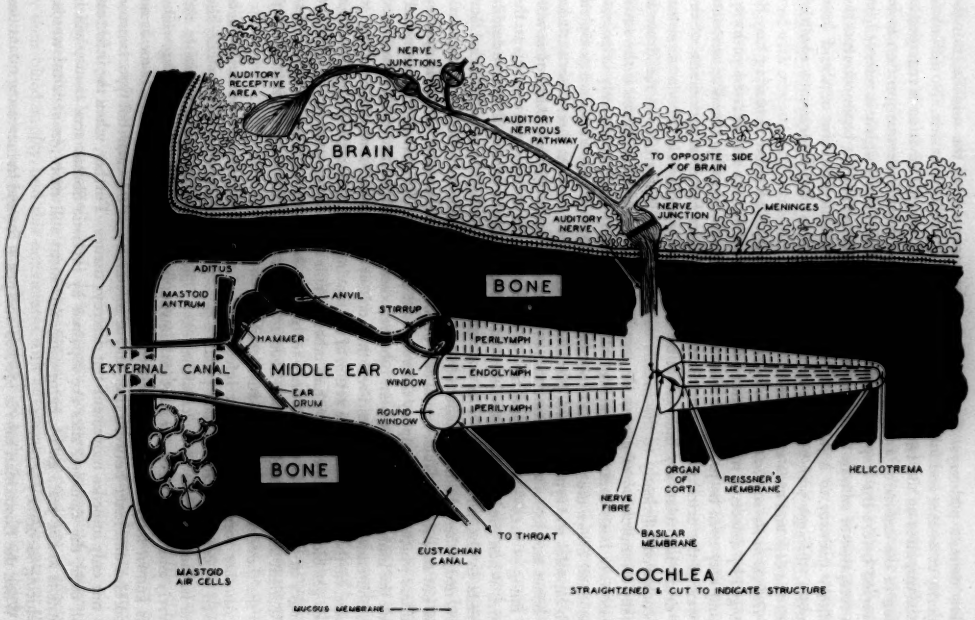


Figure 1. Diagrammatic sketch to illustrate structure of ear.

follow definite pathways. Some keep on the same side of the midline, but the majority cross over and follow a symmetrical course on the other side. No difference in the impulses on the two sides has been detected, and the division is regarded as being in the nature of a safety device. There is a second nervous junction to be traversed before the impulse reaches that part of the brain which is reserved for the reception of stimuli from the ears. A side track within the brain connects the auditory pathway to a junction which receives stimuli from the eyes and from muscles, through which co-ordination of movements with impressions obtained from the ears and eyes are effected.

Defects in the conduction of sound to the inner ear. The sense of hearing may be changed by defects at many parts of the complex mechanism and nervous pathway. Sound may even be prevented from reaching the ear drum by excessive amounts of wax in the external passage. Contrary to popular belief, a perforation of the ear drum has relatively little effect on the hearing; apparently air vibration can be conveyed efficiently to the middle ear without it. But a hole in the drum is usually an indication of present or past disease of the middle ear, and if the small bones have been damaged hearing will be moderately but not severely impaired. The small bones are connected together by miniature joints, which are complete with capsules and synovial fluid, just like the knee or elbow. On their outer surfaces the bones and joints are covered with mucous membrane which is continuous with the lining of the middle ear and that of the Eustachian tube connecting this to the throat. When the latter becomes infected and inflamed the trouble therefore readily spreads to the middle ear. The mucous membrane may become catarrhal and thickened, when the movements of the ear drum and the tiny joints are impeded, as is not surprising. Such restriction of movement is the cause of the temporary deafness which may accompany a severe cold in the head. The situation is worsened when the lining of the Eustachian tube becomes so much thickened that the lumen gets obliterated. Catarrhal fluid will then collect in the middle ear and be unable to escape, further preventing free action of the moving parts. Some people's Eustachian tubes get blocked more easily than others. With an ordinary cold in the head the changes in the middle ear are only temporary, but in some cases a more or less permanent state of mild infection occurs (known as chronic catarrhal otitis media) and the movements are always restricted.

Another form of middle ear infection, and a very common one in children, is caused by certain bacteria which produce pus. The route of infection is usually the same in these cases as in the catarrhal ones. The trouble may occur suddenly and severely and be soon over, or start insidiously, proceed slowly and last a long while. It is known medically as suppurative otitis media, acute or chronic. There is an important difference between this kind of infection and the catarrhal type, because pus not only impedes movement but destroys tissues. Further, when the infection is all over, the damaged sites are repaired by inelastic fibrous scar tissue, which contracts as the years go by, hampering movement still more; in medical words, producing secondary sclerosis.

When the Eustachian tube is blocked and pus is formed in increasing quantities

in the middle ear, the pressure increases, and earache results. An exit is urgently needed for the destructive fluid. A doctor will provide one by puncturing the bulging ear drum to let the pus escape. The ear drum may burst spontaneously on account of the pressure if this is not done. The hole thus accidentally made may be adequate for drainage, so that the symptoms subside; but it may be too small and get blocked up again. Pus from the middle ear can go through the narrow passage above the drum to the space in the bone behind the ear called the mastoid antrum, and drop down into the air cells in the cavity of the mastoid bone. This trap is lined with mucous membrane continuous with that of the middle ear and the throat. The only aperture is at the top and the narrow connexion between the mastoid antrum and the middle ear can be blocked all too easily when the lining membrane becomes thickened through inflammation. This is not at all a desirable occurrence, because the lining of the air cells may become destroyed and the bone itself may become infected and eroded, so that the pus may reach an important blood vessel which collects blood from the brain. In an extreme case the blood in this vein may clot; or an abscess may even form in the brain. It is to prevent this catastrophic series of events that a surgeon may intervene, and perform a mastoid operation. He will open up the interstices of the air cells, and destroy the wall between them and the middle ear cavity, thus improving the drainage. The rest is then left to nature. It should be noted that the small bones are not touched, and it is possible to have perfect hearing after a mastoid operation, if this is skilfully performed at the right time. But under some circumstances the operation is delayed until the small bones have become affected beyond repair, and hearing may then be impaired permanently. In other cases the damage to the small bones is so severe or long standing that the surgeon may decide to perform a radical mastoid operation. The contents of the middle ear are removed altogether in this procedure, excepting the stirrup bone which is fixed to the oval window of the inner ear. This improves the drainage still more than the simpler operation, and prevents the possibility of ear discharge continuing on account of the presence of an odd piece of dead bone.

In virulent or long standing infections pus may even penetrate from the middle ear through the windows communicating with the inner ear, and affect the cochlea, especially at the base. On this account one of the commonest causes of otitis interna is otitis media, and hence no assessment of the damage to the hearing can be made from the medical history alone.

There are other causes of deafness due to hardening of the tissues about which little is known. The abnormal changes may become noticeable when a person is in the early twenties, and proceed slowly throughout the remainder of life. This type, known as otosclerosis, sometimes runs in families, and is wont to affect the movement of the foot of the stirrup bone and the oval window. It may extend into the inner ear.

Defects in the reception of sound. It is scarcely necessary to remind readers of the extreme delicacy of the tissues of the internal ear. Their condition is affected by their blood supply, and this deteriorates with advancing years, the parts at the base of the cochlea, associated with the detection of high tones, being the first

to suffer. The upper limit of the auditory pitch range is on this account gradually lowered throughout life. The present Lord Rayleigh in his biography of his father tells how he first collaborated with him in experimental work as a baby, when his father made him cry by blowing a whistle of such a high pitch that none of the adults around could hear it. This upper tone deficiency, like other age changes in the body, proceeds at various rates in different people, so that in the later decades of their lives some are unable to hear bats, others doorbells, some complain that actors mumble, while others have difficulty in following conversation round a dinner table.

Disease of the cochlea which is secondary to changes in the middle ear has been mentioned in the preceding section.

Defects of the mechanisms of the middle and inner ears occur not only on account of disease, but also because of injury and of malformations. These peculiarities, which are sometimes inherited but by no means always so, occur more often in the inner ear than in the middle ear. Their nature can be more easily determined in the latter case, absence of small bones or of an opening into the outer ear being occasionally found. In the inner ear the nature of the defect can never be ascertained during life, and rarely even in post-mortem examination, as the delicate tissues within the hard bony skull deteriorate so soon that it is impossible for a pathologist to say what abnormal conditions were present during life. Observations on animals suggest that the Organ of Corti may often be underdeveloped. Many of these congenitally deaf people have some hearing for loud tones of a pitch below 2000 c./sec. It is possible that they appreciate it by a volley mechanism, as described below, the endings of the nerve fibres being stimulated, even though the hair cells are imperfect.

Defects in the middle ear are sometimes said to cause conduction deafness and of the inner ear perception deafness. A less suitable but common term for the latter is nerve deafness, which is often loosely used for any condition due to cochlear or auditory nerve damage. This ambiguity is particularly unfortunate as the outlook with regard to hearing aids differs greatly in these two cases. When the auditory nerve is really damaged, magnification of sound designed to reach the Organ of Corti, however perfect, will be futile; whereas in many cases of internal ear damage increasing the loudness of the incident sound is quite successful in procuring satisfactory hearing. There are, however, cases of internal ear damage when the Organ of Corti is so affected that its reception characteristics as well as its threshold for sound is changed, as shown by the following examples. Musical people with senile deafness sometimes find that the high pitch notes of an orchestra are inaudible to them, while those they can just hear sound out of tune. They always hear the violoncellos beautifully. People with internal ear deafness which has proceeded to different degrees in the two ears will sometimes remark that a note of the same frequency from an audiometer sounds different in quality in the two ears and may suffer mental confusion from binaural hearing. A music mistress with inner ear deafness tested with tones in octave steps from 64 to 8192 c./sec. said that the first two steps were octaves, the next interval was about a musical fifth, and

the top notes sounded like no musical note at all, although she could hear them, and gave a threshold reading reliably for them.

Although a fault in the ear is the commonest cause of deafness, this may also be due to disease of the nerve, as already indicated, or to injury of it, as in some fractures of the skull. The nervous impulses may also be interrupted on their way in the brain, or the auditory reception area itself may be affected by tumours, or by haemorrhages. These serious occurrences are fortunately comparatively rare.

Theory of reception of sound. What happens when sound reaches the inner ear is even now largely a mystery. It is still regarded as probable that there is a resonating mechanism in use for the detection and differentiation of high tones; but for low tones there is another possibility as an alternative, or as an addition. The electrical changes of potential in the auditory nerve have been shown to be identical in frequency with the vibrations of the incident sound when this is less than 1000 c./sec., and to bear a simple relation to it up to 3000 c./sec. Hence it would be possible for the auditory receptive area of the brain to know the nature of low sounds by the number of the impulses coming along the nerve, as well as, or instead of, recognizing which nervous fibre was bringing the message.

When sound of any frequency within the auditory range arrives at the cochlea, a change in electrical potential occurs there which has come to be known as the "electric response of the cochlea". This has been shown to be quite distinct from the change of electrical potential occurring simultaneously in the auditory nerve. Integrity of the basilar membrane seems to be essential for its continuance, but its real cause and its significance are still matters of dispute. The experimental work relating to the discovery and description of the characteristics of the electric response of the cochlea, to its differentiation from the action potentials of the auditory nerve, and to the resonance-volley theory of hearing has been summarized and discussed in another review published by the author elsewhere.

§ 3. TESTS OF HEARING

The commonly used tests of hearing are essentially estimations of change of auditory threshold. They are done with tuning forks, whose decay rate may or may not be known; or with a pure tone audiometer, which is essentially an oscillator and attenuator. In the first case comparison is made between the auditory threshold of the deaf person and the person making the test, in the second case with a normal line. In practice this normal line is usually copied from one instrument to another, and is based on a determination in America of the acuity of hearing of 72 apparently normal ears tested in a sound-proof room. The normal variation of hearing at different pitches should ideally be represented graphically by an area depicting the range of threshold found in a large number of young people without ear disease.

The threshold may be determined for air conduction and for bone conduction. In air conduction tests with tuning forks the fork is struck and held close to the external auditory opening of subject and observer in turn, and the time during which the observer can hear it after the subject is unable to, is measured. With a pure tone

audiometer a telephone receiver is used, of such a type that the vibration at the periphery is too small to be transmitted to the inner ear through the flesh and bone.

In bone conduction tests the foot of the tuning fork, or a special small flat receiver attached to the pure tone audiometer, is placed on the flesh over the mastoid bone behind the ear. It is well known that sound can be transmitted to the inner ear in such a way, even longer osseous paths being possible. For example, a tuning fork can be heard if placed on the forehead, or on the teeth. It is assumed that the vibration reaches the inner ear from the bone which surrounds it without traversing the middle ear. If a person has normal hearing, a sound of any frequency must be 30 to 50 db. louder for it to be perceived by bone conduction than by air conduction. How much of this difference is due to difficulty of transmission through the bone in the former case, and how much to the good conduction of the middle ear in the latter is undetermined. But to counteract a popular misapprehension it must be emphasized that no one, deaf or otherwise, can hear nearly as well by bone conduction as a normal person can hear in the usual way, that is by air conduction.

It is much more difficult to obtain accurate results in bone conduction tests than in air conduction tests. Some people hear equally well all over the mastoid bone, others hear much better at one particular place. This good spot is said to be over the space above the air cells, known as the mastoid antrum, and to be due to a resonance effect.

If a person with normal hearing listens to a fading note by bone conduction until it is just inaudible, and then puts his finger in his ear to block up the opening, the tone will again be audible. This phenomenon does not appear if the experiment is carried out in a silent room, and is said to be due to a masking effect of an active middle ear on the reception of sound by the inner ear. Why this should happen is unknown, but the observation explains why deaf people with purely middle ear damage when tested in a noisy room can hear better by bone conduction than can a normal observer by bone conduction in the same environment. A bone conduction test conducted in a sound-proof room, or with the opening of the external ear blocked, is referred to as an absolute bone conduction test, while one performed without such precautions is known as a relative bone conduction test.

In spite of the vagueness which still surrounds the theory of hearing by bone conduction, determination of its threshold is an invaluable addition to the knowledge of the air conduction threshold, because comparison of the two gives information as to whether the deviation from normal is due to changes in the middle ear or to defects in the inner ear or beyond this. Neither test will, however, provide information as to the cause of the changes. Some idea of this can usually be obtained from the medical history, and from an examination of the ear drum and Eustachian tube.

An important criticism of such tests, valuable though they undoubtedly are, is that people rarely want to listen to sounds that are barely audible, and that determinations of changes in auditory threshold give little indication as to how louder sounds will be affected. A few authors have recently become aware of the

importance of this consideration, but determinations of equal-loudness curves on people with defective hearing have not yet been done except in isolated cases.

Before leaving the subject of hearing tests, mention must be made of the gramophone audiometer. This valuable instrument was designed to pick out easily and quickly for further examination schoolchildren with slight or unilateral deafness who might otherwise be neglected. By early detection and treatment it is hoped to prevent deafness of a disabling extent in later life. The instrument consists essentially of a gramophone with an electro-magnetic pickup and 40 single headphones. A special gramophone record is used which consists of spoken numbers in groups of 3, in intensities diminishing in steps of 3 db. The faintest figures are just understandable to persons with normal hearing in a fairly quiet room. Further details of the instrument and its use can be found in the Board of Education *Report on Children with Defective Hearing*[1].

§ 4. HEARING AIDS

Resonators and collectors. There are only three types of instrument in use. The familiar kind, such as speaking tubes and trumpets, which collect sound and magnify it to some slight extent by resonance, are still useful. A convenient modification, because it leaves the hands free, is that form known as auricles. These are small horns, bent round to fit against the side of the head, with adjustable earpieces, which go into the opening of the ear. They are very useful for slightly deaf schoolchildren, for ladies who can hide them under their hair, and for some people who, not being very deaf, are by temperament incapable of coping with wires and batteries. Some auricles are said to amplify by about 15 db. in the upper frequencies, but this varies with the shape and size of the opening.

Microtelephones. Of the two electrical types of instruments, the microtelephones are the older. They consist essentially of a carbon microphone with granules, a small light earpiece, a switch and a pocket dry battery. They are sold under many names, and vary chiefly in the quality of the carbon microphone, as well as in the number and size of these. Some have in addition a simple amplifier and a variable resistance.

The disc earpiece may be replaced by a vibrating rectangular plaque which can be kept in position against the mastoid bone behind the ear by means of a metal headband. These bone conductors may also be fitted to valve amplifiers, but this is less commonly done.

Microtelephones are most suitable for cases of pure middle ear disease, on account of the extent and the nature of the acoustic changes in this condition, when the change of auditory threshold is rarely more than 50 db. and is maximal at 1000 to 2000 c./sec. It has been found unnecessary to employ instruments which magnify so much that very soft sounds can be heard. If sounds of 30–40 phons can be appreciated, it is sufficient for practical purposes. Amplification measurements on microtelephones, and on other instruments, have been made by T. S. Littler[4], and by the National Physical Laboratory for a few commercial firms. The theoretical and practical effect on the hearing of a patient with middle ear disease may be that, with a microtelephone, conversation at 50 or 60 phons becomes

possible instead of difficult. Some such instruments in which the microphone is rigidly attached to the battery, so that the latter can be used as a handle, are employed profitably by a few people who are far deafer, if they place the microphone within a few inches of the speaker's mouth. This is of practical, but limited value, as most people who are as deaf as this have changes in the inner ear. They are much more likely to need magnification of higher pitches than the microtelephone is capable of providing. Hence they hear distorted speech sounds, and while most severely deaf people trying microtelephones complain that they "can hear but cannot understand", a few have an extraordinary ability to guess usefully from imperfectly heard sounds. The sound amplification curve of a microtelephone can be varied slightly by physical changes in the carbon, and in the earpiece, but the changes which are possible are quite inadequate in extent to match different types of defective hearing.

The advantage of the microtelephone is its portability; its disadvantages, the noises inherent in the carbon microphone (which are particularly troublesome when the person is moving), the fact that it will not pick up distant sounds, and the expense of battery upkeep.

Valve amplifiers. The most adjustable type of hearing aid is the valve amplifier. These have become a reasonable equipment for a deaf person to carry since the introduction of midget valves, and are becoming increasingly so since batteries are becoming smaller and better. Another advance would come if these could be made lighter.

These instruments are usually 2- or 3-stage amplifiers, with carbon or piezo-electric microphones, a switch and volume control and a small light earpiece. The details, though ingenious in their attempt to conserve space and minimize weight, have nothing very outstandingly novel from a physical point of view. Such an instrument can now be obtained in a space of roughly 60 cu. in., and weighing 2 lb. It may be in a single container or in several, so that it can be disposed about the person of a man, if his pockets are of ample size and he is not very particular about changes in his apparent shape.

Some instruments, but not all, have incorporated a method of tone control by means of which high, middle or low tones can be preferentially magnified. Most of these use devices similar to those used in radio work, but in some cases the adjustment is made entirely by altering the characteristics of the earphone. These controls are certainly useful, are sometimes essential, but are often inadequate. For example, the extreme positions of a good tone control may give a maximum difference of 20 db. in the amplification at the frequencies of 200 and 4000 cycles; but it is by no means uncommon for there to be a difference of 40 db. between the auditory threshold of a deaf person at these two frequencies.

§ 5. PRESCRIPTION OF HEARING AIDS

It cannot be said truthfully either that precise prescription of hearing aids is yet possible, or that a universal type of hearing aid is a practical proposition.

With regard to a hearing aid many considerations both human and scientific must be taken into account, apart from the problems caused by those deaf people whose temperament introduces undue complications. Occupational factors often introduce serious difficulties, as the following examples taken from clinic experience may illustrate. A domestic servant cannot have any electrical aid always switched on in case the door bell rings. A light-weight jockey cannot carry a valve amplifier in order to hear the starter's instructions. It was difficult to suggest any instrument which was suitable for either a railway or a fish porter at work because both needed to hear instructions as well as handle heavy goods. A man who has a moderate degree of deafness due to middle-ear defect, and who wants to hear at lectures may need the same valve amplifier (the size of a box camera) as will another man, who can ordinarily only hear a shout in his ear on account of severe internal ear disease, to enable him to listen to his neighbour. The same instrument, with its volume tuned down, would allow the first man to hear general conversation at a dinner table, but would not be very convenient in size to be placed on the table or on his knees. A clerk who has to answer the telephone often will find a microtelephone with a bone conductor earpiece behind his ear more convenient than one with a disc earpiece which must be removed every time before he can do so. Many people if they can afford it have two different types of hearing aid, one for close range use, which is more easily worn, and another for functions in larger rooms. People with middle-ear or conduction deafness can usually hear intelligibly with most types of hearing aid, and whether they prefer resonators, microtelephones or valve amplifiers, and the last with a carbon or crystal microphone, will depend more on their circumstances, and their sensitivity to appearance and to adventitious noises and distorted sounds than to more scientific factors. People vary in this sensitivity in these as much as in other things.

There are some people with middle-ear disease who say that they hear better in a noise. It is generally believed that this is because their neighbours talk more loudly than usual to surmount the noise. But the writer agrees with the minority who are not convinced that all cases can be explained in this way. Certainly some people hear as well or better with a noisy carbon microphone than with a crystal; but these are exceptions. It is not difficult to imagine conditions of the diseased bones in the middle ear which would have the effect of increasing the force necessary to initiate their motion without affecting their action much after the inertia had been overcome.

The art and science of providing the most suitable hearing aid is more subtle in the case of people with inner ear and nerve damage. It is a matter of knowledge, experience and trial to determine whether the distortion of sounds of which they complain with an instrument (and sometimes without it) is attributable to an unsuitable hearing aid, or is a necessary consequence of their disorder. Many of these people are high-tone deaf because the base of the cochlea has been affected by one or more of the several processes described earlier. Hearing for the middle part of the auditory range (500 to 2000 c./sec.) may be as bad as the high, or nearly so, but deafness which is maximal in the 50 to 200 c./sec. range is so rare that it

can be ignored. Figure 2 illustrates the variety of acoustical picture which may be covered by the term "high-tone deaf".

It must be stressed again that the indication of a threshold at a given frequency is no indication as to the quality of the perception. In figure 2 the air conduction curves alone have been given for clarity, but many cases of high-tone deafness are composed of middle-ear and inner-ear factors, and no estimate of these components

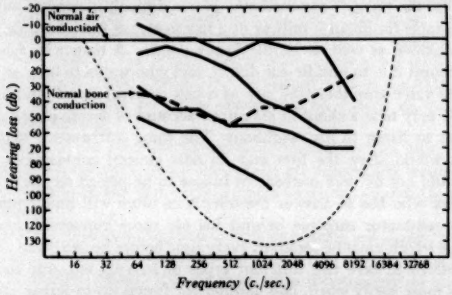


Figure 2. Auditory thresholds of four high-tone deaf persons.

can be obtained unless bone conduction thresholds are examined. High-tone deafness can occur with normal bone conduction, and be due to a pure middle-ear lesion, though this is very unusual. High-tone deafness in a person who has had middle-ear disease is much more often due to an extension of the disease to the inner ear and to be associated with an impaired bone conduction threshold for the high tones.

Few people with inner-ear damage can hear satisfactorily with microtelephones unless their ability to discriminate distorted speech sounds is unusually good. A valve amplifier in some form is essential both because of the greater power and their greater acoustical adaptability.

§ 6. TESTS WITH HEARING AIDS

The medical history and examination, the auditory threshold for air and bone conduction, and a knowledge of the deaf person's circumstances and requirements, will form a valuable guide to an experienced adviser as to the type of hearing aid which is most likely to suit him, and the manner in which this should be adjusted for him, when tone control is possible. Volume control will be regulated by the patient himself, as necessitated by the loudness of the speaker's voice, and the distance between them. Nevertheless the final test will be the practical determination of how much the deaf person can hear with the instrument. Now this is not as easy to assess as might be supposed. There are two reasons for this. Firstly, the deaf person becomes more accustomed to the instrument as time goes on. It is a common clinic experience for a deaf person with severe inner-ear disease to

say that with a valve amplifier adjusted as well as possible (though usually not according to the scientific ideal), he cannot understand a word that is being said. If the instrument is left on, and the adviser goes on talking to some one else in the room, the deaf person is often joining in the conversation within ten minutes. This phenomenon has been abused sometimes in commercial circumstances, by telling a deaf person that he will in a time, too long to be allowed for a home trial period, hear well with an instrument with which at the moment he hears nothing. In most cases the improvement in the mental appreciation of imperfectly heard sounds occurs quickly, and a home trial of a week is in most cases satisfactory, and a trial of some such period should be always recommended. Any scientific test of hearing ability with an instrument in use for the first time must therefore be of a preliminary character, although a better test can be made later profitably.

The question of a suitable test is by no means simple. It is possible to use a series of nonsense syllables pronounced by a trained speaker, as used by the Post Office Research Laboratory for testing telephone apparatus. A simplified version designed to estimate the hearing of deaf people for vowels and consonants with and without hearing aids has been described and is in use by Dr and Mrs Ewing at Manchester[2].

There is, however, another aspect to be considered. Experience with people with normal hearing and the telephone has made it plain that conversation can be understood quite well even when the acoustic transmission of it is imperfect, although difficulty with unfamiliar proper names may be experienced. Experience with deaf people and imperfect hearing aids has convinced the author that the extent of a person's ability to guess from imperfectly heard sounds forms a very important practical factor, which should not be left out of account in the assessment of the amount of benefit that he is likely to obtain from a hearing aid. A test of pure hearing suffices for a slightly deaf person, or for one for whom a perfect instrument can be provided. But for many combinations of deaf person and the best available instrument, a test of hearing and guessing is desirable. For a few people the total effect of hearing, with the help of an instrument, guessing from context, and lip-reading will together give the best measure of the help they will obtain from the use of the aid.

In collaboration with Mr D. B. Fry of the Phonetics Department of University College, two sets of tests have therefore been compiled[3]. In the first of these monosyllabic nouns are used, containing among them all the common phonetic elements of the English language. Several lists are available, so that if it is necessary to test the same person several times, memory does not play any part. These series of tests are used for people who may be expected to hear reasonably well, and is the more difficult of the two tests, although not as difficult as nonsense syllables. The second test is much easier, and is meant for people whose hearing even with an instrument is very imperfect. It consists of short and rather obvious sentences, so that if a part were heard, the rest might be guessed. Each set of sentences contains every common sound, and these in roughly the proportion in which they occur in the English language. The test words or sentences are to be said, whenever possible, by a friend of the deaf person with whose voice and accent he is familiar.

§ 7. FUTURE PROGRESS

Although it is only a few years ago that the hearing of deaf people was not tested except with uncalibrated tuning forks, and the only available hearing aids were some form of hearing tube or a microtelephone, nevertheless the present position is far from satisfactory. It must be obvious to workers in scientific fields where exact measurement is the rule how much is lacking in this sphere. The need to know the behaviour of deaf ears at various intensities of sound above the auditory threshold has already been referred to. But there are insufficient data also on the instrumental side. It is the exception to be able to obtain figures of the amplification of sound by a hearing aid at various frequencies. One needs to know not only this, but the variation of the amplification with battery potential, the amount of distortion, and the behaviour of the microphone after a period of use.

Judging from the benefits already obtained by some deaf people with imperfect instruments, it seems clear that the human ear and the brain together form a very adaptable apparatus, even when the former is diseased. But until the optimum compensation for defective hearing has been worked out scientifically in a reasonable number of cases of different types and degrees, it will be impossible to forecast what amount of approximation to the ideal would suffice for most practical purposes.

REFERENCES

References in the text

(1) Board of Education. *Report of committee of inquiry into problems relating to Children with Defective Hearing* (H.M.S.O. 1938). (Contains chapters on causation of deafness, hearing tests and hearing aids.)

(2) Ewing, I. R. and Ewing, A. W. G. *The Handicap of Deafness* (London, Longmans, 1938).

(3) Kerridge, P. M. T. "Recent advances in knowledge concerning hearing and speech." *Physiol. Rev.* **18**, 59 (1938). (Contains extensive bibliography.)

Kerridge, P. M. T. and Fry, D. B. "Tests for the hearing of speech by deaf people." *Lancet* (Jan. 1939).

(4) Littler, T. S. "Hearing aids for the deaf." *J. Sci. Instrum.* **13**, 144 (1936).

Other relevant literature

American Federation of Organizations for the Hard-of-Hearing. "Report of committee appointed to survey the instrumental aids to hearing." *Volta Rev.* **29**, 1 (1927).

Cleminson, F. J. "Hearing aids in general practice." *Brit. Med. J.* **1**, 1114 (1938). (Contains photographs of current models.)

Ewing, A. W. G., Ewing, I. R. and Littler, T. S. "The use of hearing aids." *Spec. Rep. Ser. Med. Res. Coun., Lond.*, no. 219 (1937).

Kerridge, P. M. T. "Can physics help the deaf child?" *Lancet*, p. 104 (1935).

Kerridge, P. M. T. "Aids for the deaf." *Brit. Med. J.* **1**, 1314 (1935).

Kerridge, P. M. T. "Hearing aids." *Practitioner*, **135**, 641 (1935).

Kerridge, P. M. T. "Speech defects and deafness." *Speech*, **1**, no. 2 (1935).

Kerridge, P. M. T. "Hearing and speech in deaf children." *Spec. Rep. Ser. Med. Res. Coun., Lond.*, no. 221 (1937).

KERRIDGE, P. M. T. "Hearing and speech in deaf children." *Proc. Roy. Soc. Med.* **30**, 1494 (1938).
KERRIDGE, P. M. T. "The administration of a hearing aid clinic." *J. Laryngol.* **53**, 370 (1938).
KERRIDGE, P. M. T. "Recent advances in hearing aids." *Practitioner*, **141**, 625 (1938).
MORRIS, J. "Characteristics and properties of electrical deaf aids." *J. Laryngol.* **50**, 809 (1935). (Contains a list of patents.)
REGER, S. N. "Differences in loudness response of the normal and hard-of-hearing ear at intensity levels slightly above the threshold." *Ann. Otol., etc., St Louis*, **45**, 1029 (1936).
STEINBERG, J. C. and GARDNER, M. B. "The dependence of hearing impairment on sound intensity." *J. Acoust. Soc. Amer.* **9**, 11 (1937).

General books on hearing

BEATTY, R. T. *Hearing in Man and Animals* (London, Bell, 1934).
DAVIS, A. H. *Modern Acoustics* (London, Bell, 1934).
FLETCHER, HARVEY. *Speech and Hearing* (London, Macmillan, 1929).
PHYSICAL SOCIETY. *Discussion on Audition* (Camb. Univ. Press, 1931).
RAWDON-SMITH, A. *Theories of Sensation* (Camb. Univ. Press, 1938).
STEVENS, S. S. and DAVIS, H. *Hearing* (New York, Wiley & Sons; London, Chapman & Hall, 1938).
TRENDELENBURG, F. *Klänge und Geräusche* (Springer, 1935).
WEST, W. *Acoustical Engineering* (London, Pitman, 1932).

HEAT

By R. W. POWELL, B.Sc., Ph.D.

Physics Department, The National Physical Laboratory, Teddington, Middlesex

§ 1. THE EQUATION OF STATE

The year now under review has marked the hundredth anniversary of the birth of van der Waals, and for this reason it is appropriate to give some consideration to recent developments in the field in which the name of this great Dutch physicist has been so prominent. Prior to van der Waals the limitations of the perfect gas law, $pv = RT$ had already been realized, and other workers had attempted to formulate improved equations connecting the pressure, volume and absolute temperature of a real gas. The van der Waals' equation

$$(p + a/v^2)(v - b) = RT,$$

in which a and b depend respectively on the attractive forces between the molecules and the size of the latter, was based on reasonably sound assumptions for gases of moderate density. It was highly successful in this region and achieved surprisingly good qualitative results even in the neighbourhood of the critical point. Quantitatively, however, it left much to be desired at these higher densities, and many other forms for the equation of state have since been proposed. Those suggested by Berthelot, Beattie and Bridgman, and Callendar were discussed by Awbery in these *Reports* (**1**, 176), whilst a valuable chapter of Kennard's recent book on the *Kinetic Theory of Gases* is devoted to this question [1].

Many workers, following the example of Onnes [2], express the product pv as an expansion series in powers of the density (or reciprocal volume), thus

$$pv = A + B/v + C/v^2 + \ldots,$$

or of the pressure, when

$$pv = A + B_p p + C_p p^2 + \ldots.$$

The coefficients A, B,... etc. are the so-called virial coefficients and are functions of temperature. For all gases the second virial coefficient B is negative at low temperatures but becomes positive when the temperature is sufficiently high. The temperature at which this change of sign occurs and $B = 0$ is called the Boyle temperature. A series formula including terms up to the fourth power in v was suggested by Wohl [3], whilst formulae to the fifth power have recently been given by Pinter [4] and Plank [5]. Michels and his collaborators [6, 7, 8], however, have recently studied the isotherms of carbon dioxide up to 3000 atmospheres and obtained results which cannot be represented as simple power functions of density or pressure.

A paper contributed by Lennard-Jones [9] to the van der Waals Centenary Congress gives an interesting review of the subject and a new treatment for the

critical region. He states that "van der Waals's equation of state may be regarded as the first successful attempt to interpret the equation of state of a gas in terms of interatomic forces, and conversely, as the first attempt to derive information about molecular fields from the equation of state". The present revival of interest in the equation of state is due in part to the information it supplies concerning intermolecular fields. The van der Waals form results for the case in which the molecules are considered as rigid spheres surrounded by an attractive field. Lennard-Jones himself[10, 11] considered that the potential energy, ϕ, of two molecules could be more satisfactorily represented by the expression $\phi = \lambda r^{-n} + \mu r^{-m}$, i.e. as the sum of attractive and repulsive forces which vary as inverse powers of the distance, r, between the molecules. The best agreement with experimental results for the inert gases was obtained when $n = 12$ and $m = 6$. As only binary encounters between the atoms of the gas had been considered, the equations of state then obtained only held for dilute gases. Equations of state, for the particular case in which the atoms are assumed to be hard elastic spheres, have been worked out for both dilute and dense gases by Tonks[12] on the basis of the virial theorem of Clausius.

In their more general treatment for the dense gas Lennard-Jones and Devonshire[13] consider the average position of each atom to be more or less fixed as in a solid. This conception, though it neglects the possibility of diffusion, is undoubtedly much nearer the truth for a dense gas than that which only considered the effect of binary encounters. Methods of statistical mechanics are then used to evaluate the partition function for such a gas employing the same potential fields as had been previously obtained for the gas in the rarefied state. This treatment appears to be directed along the right lines, for not only is the equation of the correct kind, but the Boyle and critical temperatures of several gases are given closely, as can be seen from the data in table 1.

Table 1. Calculated and observed properties of gases

Gas	Boyle temperature (°K.)		Critical temperature (°K.)	
	Observed	Calculated	Observed	Calculated
Helium	23	23·4	5·2	9·1
Hydrogen	109	105	33·1	40·9
Neon	123	121	44·7	47·1
Nitrogen	323	328	126	128
Argon	410	408	149·7	159·3
Oxygen	423	—	155	(164)
Carbon monoxide	333	—	134	(130)
Methane	491	—	190·5	(191)

In this table the observed Boyle temperatures have been used in deriving the values enclosed in brackets. Values of 26·7°, 85·8° and 71·8° K. have been calculated for the respective boiling points of neon, argon and nitrogen, which compare reasonably well with observed values of 27·2°, 87·4° and 77·2° K.

Devonshire[14] has since used the same method to determine the equation of state of a two-dimensional gas, a case also considered by Tonks [12]. For such a

gas he finds the critical temperature to be only 53 per cent of that of the three-dimensional gas previously considered. The nearest approach to a two-dimensional gas to be studied experimentally is an inert gas adsorbed on a surface of a liquid. Cassel and Neugebauer[15] had been unsuccessful when they endeavoured to detect critical phenomena for krypton and xenon when adsorbed on liquid mercury. Devonshire's result suggests that these investigations might prove successful if extended to much lower temperatures.

An attempt to obtain a rigorous derivation of the equation of state from the standpoint of statistical mechanics for any law of force between the molecules has recently been made by Mayer and his colleague[16, 17]. A system composed of identical molecules having mutual potential energy is considered, and the assumption made that the total potential energy can be expressed as the sum of that between all the pairs of molecules. This allows the derivation of accurate formal equations for the thermodynamic properties of the system. Under conditions generally obtained at low temperatures a region is indicated where the pressure and Gibbs free energy become independent of volume, and condensation to the liquid phase can be understood to occur. Born[18] was sufficiently impressed with this new method of approach to bring a paper based on Mayer's work to the notice of the above-mentioned Centenary Congress, and later, in collaboration with Fuchs[19], he has himself contributed a further paper on this subject. This latter treatment predicts the existence of triple points as well as a critical point.

The method is still very much in the development stage, and Kahn and Uhlenbeck[20] have also contributed a paper on this theory of condensation in which they point out that some of Mayer's reasoning and the equations derived are analogous to those of Einstein for the ideal Bose gas.

The analysis has since been carried a stage further by Mayer and Harrison[21, 22], who have shown that their equations predict the existence of some unexpected phenomena at the critical point. At a temperature T_m, a little below the critical temperature T_c, the surface tension of the condensed phase is zero, and this phase undergoes a discontinuous change; between T_m and T_c there is a finite volume interval for which pressure and the Gibbs free energy are independent of volume. A theoretical result of this nature is of considerable interest, for experiments by Callendar[23] had indicated that for water there was a "critical region" extending for some 6° C., as the dividing line between liquid and vapour persisted to this extent above the accepted critical temperature of 374° C., and the densities in the sections where the two phases had existed differed appreciably. Maass and his collaborators[24, 25, 26] have also described experiments indicating the existence of complicated phenomena in the critical region.

References

(1) KENNARD, E. H. *Kinetic Theory of Gases*, p. 206 (McGraw Hill Publ. Co. 1938).
(2) ONNES, K. *Commun. Phys. Lab. Leiden*, No. 71 (1901); *Proc. K. Akad. wet. Amst.* **4**, 125 (1900).
(3) WOHL, A. *Z. phys. Chem.* **99**, 207 (1921).

(4) PINTER, T. *Acta phys. polon.* **4**, 23 (1935).
(5) PLANK, R. *ForschArb. IngWes.* **7**, 161 (1936).
(6) MICHELS, A. and MICHELS, C. *Proc. Roy. Soc.* A, **160**, 348 (1937).
(7) MICHELS, A., BIJL, A. and MICHELS, C. *Proc. Roy. Soc.* A, **160**, 376 (1937).
(8) MICHELS, A., BLAISSE, B. and MICHELS, C. *Proc. Roy. Soc.* A, **160**, 358 (1937).
(9) LENNARD-JONES, J. E. *Physica*, **4**, 941 (1937).
(10) LENNARD-JONES, J. E. *Proc. Roy. Soc.* A, **106**, 463 (1924).
(11) LENNARD-JONES, J. E. *Proc. Phys. Soc.* **43**, 461 (1931).
(12) TONKS, L. *Phys. Rev.* **50**, 955 (1936).
(13) LENNARD-JONES, J. E. and DEVONSHIRE, A. F. *Proc. Roy. Soc.* A, **163**, 53 (1937).
(14) DEVONSHIRE, A. F. *Proc. Roy. Soc.* A, **163**, 132 (1937).
(15) CASSEL, H. M. and NEUGEBAUER, K. *J. Phys. Chem.* **40**, 523 (1936).
(16) MAYER, J. E. *J. Chem. Phys.* **5**, 67 (1937).
(17) MAYER, J. E. and ACKERMANN, P. G. *J. Chem. Phys.* **5**, 74 (1937).
(18) BORN, M. *Physica*, **4**, 1034 (1937).
(19) BORN, M. and FUCHS, K. *Proc. Roy. Soc.* A, **166**, 391 (1938).
(20) KAHN, B. and UHLENBECK, G. E. *Physica*, **5**, 399 (1938).
(21) MAYER, J. E. and HARRISON, S. F. *J. Chem. Phys.* **6**, 87 (1938).
(22) HARRISON, S. F. and MAYER, J. E. *J. Chem. Phys.* **6**, 101 (1938).
(23) CALLENDAR, H. L. *Proc. Roy. Soc.* A, **120**, 460 (1928); *Engineering*, **126**, 594 (1928).
(24) TAPP, J. S., STEACIE, E. W. R. and MAASS, O. *Canad. J. Res.* **9**, 217 (1933).
(25) WINKLER, C. A. and MAASS, O. *Canad. J. Res.* **9**, 613 (1933).
(26) MAASS, O. and GEDDES, A. L. *Philos. Trans.* A, **236**, 303 (1937).

§ 2. THERMAL DIFFUSION

Thermal diffusion is another phenomenon which throws considerable light on the nature of molecular force fields. The possibility of the existence of this effect was predicted independently by Chapman[27] and Enskog[28]; the experimental evidence for it was obtained by Chapman and Dootson[29]. The theory indicated that when a temperature gradient is applied to a uniform mixture of two gases separation should occur, the larger and heavier molecules moving towards the cold side and the lighter and smaller ones to the hot side. This separating effect is balanced by the mixing effect of ordinary gaseous diffusion so that finally a steady gradient of concentration is set up by that of temperature. The theory shows that for a given temperature difference the amount of thermal separation depends on the ratios of the masses and diameters of the molecules, on the original proportions by volume of the gases and, also, on the nature of the force-field between the unlike molecules. The effect is greatest for molecules which behave as hard elastic spheres and disappears entirely for molecules which obey the inverse fifth power law of repulsion assumed by Maxwell. Chapman and Dootson's experiments, using mixtures of hydrogen and carbon dioxide and hydrogen and sulphur dioxide, showed that these molecules were not of this Maxwellian type, as a separation was observed, which was rather less than half of the anticipated separation for hard elastic spheres.

Many experimental investigations of the magnitude of thermal diffusion for various pairs of gases have since been carried out, notably by Ibbs[30,31], Lugg[32], Grew and Atkins[33] and G. and O. Blüh[34]. The first-mentioned worker has recently published[35] a comprehensive account of the experimental development of the subject. Least separation is observed for nitrogen-carbon dioxide mixtures and

most for helium-argon mixtures. Molecules, such as those of helium, which invariably cause large separation are termed hard molecules, whilst soft molecules are those which tend to obey the inverse fifth power law. Ibbs and Wakeham[36] find that the molecular field of carbon dioxide changes from soft to that of a much harder molecule at about 145° c., though Lugg[32] regards this change as being a gradual one.

Chapman[37] subsequently published a special treatment for mixtures of two gases in which one component is present in very small quantity. For this case the force index between the unlike molecules can be simply determined from a measurement of the ratio of the proportion by volume of the dilute constituent on the cold side to that on the hot side, together with a knowledge of the coefficients of ordinary diffusion of the two gases and the viscosity of the lighter gas. Harrison[38] has recently used this result to study the thermal diffusion of radon in hydrogen and in helium. The values he obtains for the repulsive force indices of radon in these two gases are 6·3 and 7·3 respectively. These low values indicate that the radon molecule must be very "soft", as hydrogen and helium are both known to be very hard molecules. This is the first information obtained regarding the field of this molecule of high molecular weight. Puschner[39] has recently obtained values of 10 and 15 for the repulsive force indices of helium with neon and argon respectively, and his values can be compared with those of 11·4 and 9·0 obtained by Ibbs and Grew[40] for the same gas mixtures.

References

(27) Chapman, S. *Philos. Trans.* A, **217**, 115 (1918).
(28) Enskog, D. Dissertation, Upsala (1917).
(29) Chapman, S. and Dootson, F. W. *Phil. Mag.* **33**, 268 (1917).
(30) Ibbs, T. L. *Proc. Roy. Soc.* A, **99**, 385 (1921).
(31) Ibbs, T. L. *Proc. Roy. Soc.* A, **107**, 470 (1925).
(32) Lugg, J. W. H. *Phil. Mag.* **8**, 109 (1929).
(33) Grew, K. E. and Atkins, B. E. *Proc. Phys. Soc.* **48**, 415 (1936).
(34) Blüh, G. and O. *Z. Phys.* **90**, 12 (1934).
(35) Ibbs, T. L. *Physica*, **4**, 1133 (1937).
(36) Ibbs, T. L. and Wakeham, A. C. R. *Proc. Roy. Soc.* A, **134**, 613 (1932).
(37) Chapman, S. *Phil. Mag.* **7**, 1 (1929).
(38) Harrison, G. E. *Proc. Roy. Soc.* A, **161**, 80 (1937).
(39) Puschner, M. *Z. Phys.* **106**, 597 (1937).
(40) Ibbs, T. L. and Grew, K. E. *Proc. Phys. Soc.* **43**, 142 (1931).

§ 3. THE MEASUREMENT OF TEMPERATURE

The international temperature scale has not yet been defined below −182·97° c., the boiling point of oxygen, and so it remains necessary for the various workers, interested in these low temperatures to establish their own fixed points and temperature scales. Reference was made last year to the work of Henning and Otto[41] at the Reichsanstalt. The temperature scale in use at the Leiden laboratory over the liquid helium range was settled in 1932[42] and was based on measurements carried

out three years earlier[43, 44]. Since that time experience in low temperature measurements and the knowledge of the laws of rarefied gases have considerably advanced and, accordingly, a new series of measurements have been undertaken by Schmidt and Keesom[45, 46] with a view to fixing the scale more definitely over this range. A helium thermometer with reduced ice point was used, pressures being read by a hot wire anemometer. With this arrangement it was hoped to achieve the same order of accuracy as that obtainable at higher temperatures with the normal helium thermometer. They obtain a value of 4·216° K. for the boiling point of helium and 2·186° ± 0·002° K. for the lambda point. The temperatures throughout the range are slightly higher than calculated according to the 1932 scale, the maximum differences amounting to 0·026° for helium II at 1·6° K. and 0·015° for helium I at 3·5° K. These data are based on a value of 20·42° K. for the boiling point of hydrogen, whereas Henning and Otto[41] attributed a value of 20·380° K. to this point.

Further attempts have been made to establish satisfactory experimental methods for the measurement of temperatures below 1° K. The first method to be used was based on a measurement of magnetic susceptibility from observation of the ballistic throw produced by a small search-coil. This method involves the assumption that the salt obeys Curie's law, and temperatures measured according to this Curie scale are denoted by T^*. Allen and Shire[47] claim that the same order of accuracy can be obtained over the range 0·028°* to 0·10°* if a resistance thermometer of fine phosphor-bronze wire is used. The calibration depends on the current flowing in the wire, but provided this current is kept the same and measurements are made in the absence of any external magnetic field the wire can be subjected to strong magnetic fields without altering the temperature indication of the instrument. For use in very strong magnetic fields van Laer and Keesom[48] recommend the addition of upwards of 0·3 per cent of a lead-bismuth alloy to the phosphor-bronze. Giauque, Stout and Clark[49] have suggested the use of amorphous carbon as a resistance thermometer for measuring low temperatures.

The preliminary experiments of Giauque and MacDougall[50] to ascertain how the Curie scale differs from the thermodynamic scale of temperature have been followed by experiments by Kürti, Lainé and Simon[51]. These workers find that Curie's law does not hold for iron ammonium alum below 0·1° K. They note a maximum departure at 0·06° K. and also a sudden bend in the susceptibility curve at 0·034° K., which is thought to be the Curie point for the salt. The matter has been investigated theoretically by Hebb and Purcell[52], on the assumption that the departure is due to the electric crystalline fields acting on the magnetic ions of the paramagnetic salts and the magnetic dipole-dipole coupling between these ions. Numerical calculations of the specific heat and susceptibility of several of the salts used have been made, and tables prepared showing the anticipated percentage difference between T^* and T for these salts. According to these calculations the lowest temperature yet obtained[53], namely 0·0043°*, reduces to 0·0034° K.

At the other end of the temperature scale it is seen that Hoffmann and Schulze[54] have redetermined the melting point of silicon. They conclude this

temperature to be 1410° ± 2° C., whereas earlier values had ranged from 1404° to 1430° C. Hoffmann and Tingwaldt[55] consider that a carbon arc can be used as a pyrometric standard. Sighting with a disappearing filament pyrometer on the positive crater of an arc with carbon and graphite anode they obtain a reproducible value of 3840° K., which is in close agreement with earlier observations.

Schofield[56] has published his first report on the measurement of liquid steel temperatures, which describes attempts to adapt the platinum thermocouple and the optical pyrometer to the conditions of the open-hearth furnace. Grüss[57] has pointed out that some varieties of refractory materials which become reasonably good electrical conductors at high temperatures can then be used as resistance thermometers.

A modification in the design of optical pyrometers of the disappearing filament type has been suggested by Stott[58], in which a slit set parallel to the filament replaces the small circular aperture between the erecting lens and the eyepiece. In this way the brightness of the field is increased and the useful range of the instrument extended by some 80° C. at the lower end of the scale.

References

(41) Henning, F. and Otto, J. *Phys. Z.* **37**, 601, 633 and 639 (1936).
(42) Keesom, W. H. *Commun. Phys. Lab. Univ. Leiden*, Suppl. no. 71*d*, (1932).
(43) Keesom, W. H., Weber, S. and Nørgaard, G. *Commun. Phys. Lab. Univ. Leiden*, no. 202*b*.
(44) Keesom, W. H., Weber, S. and Schmidt, G. *Commun. Phys. Lab. Univ. Leiden*, no. 202*c*.
(45) Schmidt, G. and Keesom, W. H. *Commun. Phys. Lab. Univ. Leiden*, no. 250*b*; *Physica*, **4**, 963 (1937).
(46) Schmidt, G. and Keesom, W. H. *Commun. Phys. Lab. Univ. Leiden*, no. 250*c*; *Physica*, **4**, 971 (1937).
(47) Allen, J. F. and Shire, E. S. *Nature, Lond.*, **139**, 878 (1937).
(48) van Laer, P. H. and Keesom, W. H. *Commun. Phys. Lab. Univ. Leiden*, no. 252*f*; *Physica*, **5**, 541 (1938).
(49) Giauque, W. F., Stout, J. W. and Clark, C. W. *J. Amer. Chem. Soc.* **60**, 1053 (1938).
(50) Giauque, W. F. and MacDougall, D. P. *Phys. Rev.* **47**, 885 (1935).
(51) Kürti, N., Lainé, P. and Simon, F. *C.R. Acad. Sci., Paris*, **204**, 754 (1937).
(52) Hebb, M. H. and Purcell, E. M. *J. Chem. Phys.* **5**, 338 (1937).
(53) de Haas, W. J. and Wiersma, E. C. *Physica*, **2**, 335 (1935).
(54) Hoffmann, F. and Schulze, A. *Phys. Z.* **38**, 901 (1937).
(55) Hoffmann, F. and Tingwaldt, C. *Wiss. Abh. phys.-tech. Reichsanst., Berl.*, **21**, pt. 4, 38 (1937).
(56) Schofield, F. H. *Seventh Report on the Heterogeneity of Steel Ingots, Iron and Steel Institute* p. 215 (1937).
(57) Grüss, H. *Arch. tech. Messen*, no. 78, T. 161 (1937).
(58) Stott, V. H. *J. Sci. Instrum.* **14**, 370 (1937).

§4. CALORIMETRY AND SPECIFIC HEAT

In recent years the accurate determination of the specific heats of certain alloys over a wide range of temperature has become a matter of prime importance, and Sykes[59] has been the leading experimentalist in this field. This interest has arisen

since it has been realized that increase in the temperature of an alloy can cause a transformation to occur from a state of atomic order to one of disorder. For a comprehensive account of this subject of the order-disorder transformations which occur in alloys reference should be made to a paper by Nix and Shockley[60], whilst a summary of the numerous contributions of the Manchester school has been given by Bragg, Sykes and Bradley[61].

The simplest theoretical treatment is that of Bragg and Williams[62], who use a definition in which the degree of order, S, depends on the fraction of atoms which are in the positions appropriate to the completely ordered state. In such a state the atoms form what is termed a superlattice and S has the value of unity. With increase in temperature a few atoms jump into wrong positions until ultimately a certain critical temperature T_c is reached, at which the ordinary forces are overcome and the alloy becomes completely disordered. Then S has a value of zero. S is calculated as a function of temperature for alloys of the AB and AB_3 types. Energy is, of course, required to produce this disordered state, which manifests itself as an anomalous specific heat, the specific heat rising to a sharp maximum just below T_c, in a manner analogous to that of a ferromagnetic metal at the Curie point. Bethe[63] developed a theory which allowed for the existence of local ordering forces between neighbouring atoms, and showed that ordering of this type would still persist when the superlattice order had entirely vanished, and this would result in the specific heat above T_c remaining greater than the normal Dulong-Petit value. This brief account serves to show the value of accurate specific heat measurements, since not only are such measurements an important means for investigating superstructures but they serve as tests for the various theories.

The method used by Sykes[59] for measurements on β brass was described in these *Reports* (**3**, 156). The total energy involved in the transformation was then found to be about 85 per cent of that predicted by the Bragg-Williams theory. The curve above T_c remained high after the manner suggested by Bethe. Specific heat-temperature curves have since been obtained by Sykes and Wilkinson[64] for a number of α, $\alpha+\beta$ and β brasses using an improved apparatus. These curves, together with energy content measurements, enable a reliable estimate to be made as to the behaviour of a β brass having the exact composition, CuZn, considered by the theorists. The change in specific heat at the critical temperature is found to be greater than that indicated by either of the above-mentioned theories, whilst at lower temperatures the experimental curve lies below both theoretical curves. It thus seems that the disappearance of order takes place more suddenly than either theory suggests. The energy associated with the transformation below the critical temperature is found experimentally to be 9·8 cal./g., whereas 11·4 and 10·9 cal./g. are indicated by the respective theories of Bragg and Williams and of Bethe.

These experimental results are in very good agreement with a determination by Moser[65] in which a similar method was employed, whereas the value derived by Steinwehr and Schulze[66] from cooling curve data is much smaller. It seems probable that only a portion of the transformation was included in this latter case.

Easthope[67] has endeavoured to show theoretically how the critical ordering of alloys depends on their composition and his predictions for alloys of the AB type are well confirmed by the data of Sykes and Wilkinson[64]. These workers find the energy needed to transform the alloy from an highly ordered state to a disordered state (over the same temperature range) to increase as the zinc content approaches 50 per cent. It is only 75 per cent of this maximum value for 45 per cent zinc, whereas Easthope predicts 78 per cent for an alloy of this composition. There is also an accompanying decrease in T_c which agrees reasonably well with Easthope's predictions. Theory and experiment, however, appear to be less in agreement for alloys of the AB_3 type.

Mott[68] has been responsible for obtaining an independent estimate of the energy required to destroy the superlattice of β brass, and hence of the heat absorbed in the transition from the ordered to the disordered state. The electronic theory of metals is used to obtain the electrostatic energy of the superlattice (0·027 ev./atom) and the energy due to exchange repulsion between the copper and zinc ions (0·013 ev./atom). Taking the total energy per atom of the superlattice as the sum of these two quantities a value of 0·040 ev. is obtained. This estimate for the total energy appears to be of the order of magnitude indicated by the Bragg-Williams theory for this transformation, from which Mott derives an approximate value of less than 0·043 ev./atom. There is also some indication that Mott's theory is in closer accord with experiment, in that it indicates a more sudden rise in specific heat at the critical temperature.

Many crystalline solids have also been found to have a somewhat analogous anomalous specific heat, thus with increase in temperature, the specific heat of ammonium chloride increases to a sharp maximum at about $-30°$ C.[69], and Fowler[70] obtains a qualitative explanation of this by very similar reasoning to that of Bragg and Williams. The latter assume the energy required to move an atom from an ordered to a disordered position to be proportional to the average degree of order of the whole structure, whereas Fowler introduces the idea of molecular rotations, and assumes that the potential energy of the rotating molecule must be a function of the amount of rotation already present. Nitta and Suenaga[71] are seeking to throw additional light on this particular case, by experiments using ND_4Cl, in which the moment of inertia of the ammonium ion is doubled, whilst the lattice constant and the potential of the crystal remain practically unchanged. Results for a preliminary sample of the salt indicate that the peak now occurs at $-24°$ C. Bearing on this, Smits and Muller[72] reported some time ago a heterogeneous change in the volume-temperature curve of NH_4Cl at $-30°$ C. which is accompanied by hysteresis, whereas for ND_4Cl they find[73] a continuous transition and no hysteresis, although the curve is steepest at about $-24°$ C.

A vacuum calorimeter has been used by Perlick[74] for the measurement of specific heats, latent heats of transformation and heat of vaporization at temperatures ranging from $-190°$ C. to room temperature. He was mainly interested in refrigerants, and gives results for sulphur dioxide, methylene chloride, difluoro-mono-chloro-ethane and tetra-fluoro-dichloro-ethane. The specific heat-

temperature curve of this latter substance has discontinuous peaks at about $-95°$, $-140°$ and $-170°$ C.

Giauque and Egan[75] have published a detailed description of the apparatus used for many of the calorimetric researches at low temperatures, which have been undertaken at the University of California, with a view to comparing entropy values calculated by means of the third law of thermodynamics with values derived from band spectral data. The same apparatus is also employed for vapour pressure measurements on condensed gases, and this paper contains results for carbon dioxide. The heat of vaporization at the sublimation point, 194·67° K., is given as 6030 cal./mole. The heat capacity of the solid form has been measured over the range 15° K. to the sublimation point, and the values are found to agree well with those of Eucken and Hauck[76], the maximum deviation being 4·3 per cent at 120° K.

Carbon dioxide was known to be a molecule of the linear symmetrical type, yet Kelley[77] had shown that the data available at low temperatures indicated a third law entropy value about 1 unit greater than the spectroscopic value. Calculations based on this new experimental data have removed the discrepancy, the experimental and spectroscopic entropy values at 194·67° K. being respectively 47·55 and 47·59 cal./degree, and those at 298·10° K. being 51·07 and 51·11.

Schmidt[78] has published a review of existing data, which confirms this close agreement between absolute entropy values obtained by the two means mentioned above for the technically important gases, and shows that equilibrium constants obtained from either entropy differ inappreciably from those determined experimentally. The paper includes tables of specific heats and functions to be used in the calculation of equilibrium constants for temperatures ranging from 273° to 3000° K.

Before leaving the subject of specific heats mention should be made of the work of Davies and Thomas, who have worked out the theory[79] of an electrical method of determining this constant, in which the specimen is heated by eddy currents produced within it when placed in an alternating magnetic field. The method has been applied[80] to determine the specific heat at 20° C. of a number of metals. Three of these specimens were the same as had been previously examined by E. H. and E. Griffiths[81] and the two sets of values agreed to within $\frac{1}{4}$ per cent. The method can be adapted for testing liquids, but the value obtained for water is 1·5 per cent too large. The error is attributed to the difficulty of obtaining sufficient uniformity of temperature throughout the system.

References

(59) Sykes, C. *Proc. Roy. Soc.* A, **148**, 422 (1935).
(60) Nix, F. C. and Shockley, W. *Rev. Mod. Phys.* **10**, 1 (1938).
(61) Bragg, W. L., Sykes, C. and Bradley, A. J. *Proc. Phys. Soc.* **49**, (Extra Part), 96 (31 August 1937).
(62) Bragg, W. L. and Williams, E. J. *Proc. Roy. Soc.* A, **145**, 699 (1934); **151**, 540 (1935).
(63) Bethe, H. A. *Proc. Roy. Soc.* A, **150**, 552 (1935).
(64) Sykes, C. and Wilkinson, H. *J. Inst. Met.* **61**, 223 (1937).

(65) Moser, H. *Phys. Z.* **37**, 737 (1936).
(66) Steinwehr, H. and Schulze, A. *Phys. Z.* **35**, 385 (1935).
(67) Easthope, C. E. *Proc. Camb. Phil. Soc.* **33**, 502 (1937).
(68) Mott, N. F. *Proc. Phys. Soc.* **49**, 258 (1937).
(69) Simon, F. *Ann. Phys., Lpz.*, **68**, 241 (1922).
(70) Fowler, R. H. *Proc. Roy. Soc.* A, **149**, 1 (1935) and **151**, 1 (1935).
(71) Nitta, I. and Suenaga, K. *Sci. Pap. Inst. Phys. Chem. Res., Tokyo*, **32**, 8$\frac{1}{3}$ (1937).
(72) Smits, A. and Muller, G. J. *Z. phys. Chem.* A, **166**, 97 (1933).
(73) Smits, A. and Muller, G. J. *Nature, Lond.*, **139**, 804 (1937).
(74) Perlick, A. *Z. ges. Kälteindustr.* **44**, 201 (1937).
(75) Giauque, W. F. and Egan, C. J. *J. Chem. Phys.* **5**, 45 (1937).
(76) Eucken, A. and Hauck, F. *Z. phys. Chem.* **134**, 161 (1928).
(77) Kelley, K. K. *Bull. Bureau Mines*, **350**, 20 (1932) and **383**, 35 (1935).
(78) Schmidt, F. A. F. *ForschArb. IngWes.* **8**, 91 (1937).
(79) Davies, R. M. and Thomas, W. J. *Phil. Mag.* **24**, 600 (1937).
(80) Thomas, W. J. and Davies, R. M. *Phil. Mag.* **24**, 713 (1937).
(81) Griffiths, E. H. and Griffiths, E. *Philos. Trans.* A, **213**, 119 (1913).

§ 5. THE CONDUCTION OF HEAT

Heat conduction in electrical insulators. A knowledge of the thermal conductivity of the insulation used in the construction of electrical machinery is a matter of practical importance; for this insulation, which is necessary to restrict the flow of electricity to certain paths, frequently intervenes between the electrical conductor and the air or oil used as cooling medium. Much of the heat generated in the conductor has, therefore, to be dissipated through the insulation, and it follows that the performance of the machine is determined by the heat conducting properties of the electrical insulation. Efforts are made to obtain good electrical insulators of high thermal conductivity. The thermal conductivities of materials of this type are usually measured in some form of plate apparatus. A description of some improvements in the equipment used for such tests at the National Physical Laboratory has been published by Griffiths[82], whilst Weh[83] has described the method which has been in use at the General Electric Company for several years. Many new insulators which consist of synthetic resins with various types of filling materials have been introduced within the last few years, and Erk, Keller and Poltz[84] have published results for the thermal conductivity, at about 40° C., of no less than 46 of these materials. The method used was a comparative one. The absolute thermal conductivity of a sample of Borkron plate glass had been accurately determined and this was used as the standard material. Its conductivity increased from 0·00258 to 0·00275 cal./cm.-sec.-°C. over the temperature range 20° to 100° C. The standard plate was 117 mm. in diameter and 5 mm. thick. At the top of the apparatus was an aluminium plate through which hot water was circulated, and at the base a water-cooled aluminium cylinder. In between, starting from the top, were a 10 mm. thick copper plate, the test plate, a 2·5 mm. copper plate, the standard plate and another 10 mm. copper plate. Pressure was applied to the pile of plates by means of a screw and all contacts were improved by thin films of oil. The temperatures of the copper plates were measured by thermocouples, and corrections applied to allow for the temperature drop through the oil film.

An inherent disadvantage of all direct methods of measuring thermal conductivity is the length of time which has to be allowed to ensure that steady conditions are established. Dannatt[85] has recently suggested the following approximate method, by means of which, he states that the thermal conductivity at normal temperatures can be determined to within 5 per cent in under 10 min. The sample to be tested, in the form of a disc about 1·5 in. in diameter and 0·1 in. thick, is placed on a massive copper plate heated to a constant temperature. A copper disc, of the same diameter and about ¼ in. thick, is placed on the sample, and immediately on top of this is placed a calorimeter full of water at a temperature a few degrees less than that of the surroundings. Since the heat capacity of this calorimeter is also large, a temperature drop, which can be regarded as constant, is soon set up across the specimen. At a noted time the calorimeter is removed. This allows the temperature of the copper disc to increase. Its temperature-time curve is plotted from thermocouple measurements, and from this observed rate of temperature rise and a knowledge of the thermal capacity of the disc the heat flow through the test material can be determined. Knowing the temperature drop across the specimen its thermal conductivity can be derived. To obviate any correction for heat interchange with the surroundings, the conductivity should be calculated from the observed rate at which the disc passes through a temperature equal to that of the surroundings. The effect of the thermal capacity of the specimen itself can be allowed for by adding a small correction, equal to approximately one-third of the thermal capacity of the specimen, to that of the copper disc whose temperature rise is observed.

In the particular case of mica, which is used extensively in sparking plugs and in the construction of electric heaters, it is of importance to study the thermal conductivity over a much wider range of temperatures. Powell and Griffiths[86] have published results for the thermal conductivities of a number of micas up to 600° C. They used the divided bar method, a disc of mica about 1 mm. in thickness being clamped between two metal rods of known thermal conductivity, and the energy flowing across the mica being derived from the temperature gradients established in the rods. Corrections for temperature discontinuities at the mica-metal interfaces were obtained from experiments in which very thin mica specimens were used. Five specimens of Indian muscovite micas showed but slight variation of thermal conductivity with temperature over the range 100° to 600° C., all values lying within 18 per cent of a mean of 0·0016 cal./cm.-sec.-°C. A much more interesting result was obtained for certain varieties of phlogopite, which showed, in varying degree, a pronounced decrease in thermal conductivity in the neighbourhood of 200° C. A sample of Canadian phlogopite, for instance, had a thermal conductivity of about 0·0014 at 100° C.; at 250° C. it had fallen to 0·00043 and remained of this order on further heating, whilst on cooling it only partially recovered, having a value of about 0·0011 at 100° C. An X-ray examination of the same samples of mica, when heated through a temperature cycle ranging from 16° to 400° C., was subsequently carried out by Wood[87], and it was found that those samples for which this large decrease in the thermal conductivity had been obtained gave very ill-

defined x-ray spectra from 200° C. upwards. From these results it was concluded that the heat treatment may cause a very marked mosaic formation in the mica, which is not stable and is only partially reversible on cooling. It is no doubt this change in the mica structure—a change that was not apparent to the eye—which accounts for the marked decrease in thermal conductivity.

Heat conduction in non-metallic crystals. Theoretical papers on the conduction of heat in an ideal crystal were first published by Debye[88], who was able to show that the conductivity should be inversely proportional to the absolute temperature. Peierls[89] later showed the processes involved to be more complex than those considered by Debye; whereas the latter's result should hold at high temperatures, at low temperatures the conductivity should vary inversely as the square of the absolute temperature. The same result was later obtained in a more complete presentation by Papapetru[90]. Blackman[91] has considered the relative conductivities of various simple cubic crystals of the NaCl type, and has shown that we should expect the thermal conductivity to have a maximum value when the masses of the ions are equal. This result is very strikingly supported by the earlier measurements of Eucken and Kuhn[92], who at 0° C., for the crystals NaF, NaCl and NaBr, obtained thermal conductivity values of 0·0252, 0·0238 and 0·0058; for KF, KCl, KBr and KI values of 0·0169, 0·0245, 0·0087 and 0·0073; and for RbCl, RbBr and RbI values of 0·0047, 0·0091 and 0·0077. In each of the three groups the conductivity assumes a maximum value for the crystal whose constituent elements have masses lying closest together. Similar relationships were established at −190° C. Thermal conductivity measurements on KBr and KCl crystals have recently been carried to much lower temperatures by de Haas and Biermasz[93]. Single crystals of quartz had first been studied[94], and whereas at high temperatures the results obtained for this material agreed with theory in giving a conductivity varying inversely as the absolute temperature, below 10° K. the thermal conductivity was found unexpectedly to decrease rapidly in proportion to $T^{2\cdot5}$. Doubt was expressed as to whether this behaviour at low temperatures was a property of the ideal crystal, or whether the reduction in conductivity was brought about by lattice failures, such as the development of a mosaic structure. To extend the work to substances of greater purity than quartz the experiments were repeated using crystals of KBr and KCl. In each case a similar behaviour was noted[93], for both KBr and KCl the maximum in the thermal conductivity curve occurred at about 7° K., and at lower temperatures the conductivity became proportional to $T^{2\cdot5}$ for KBr and to $T^{1\cdot6}$ for KCl.

In this paper mention is made of a possible explanation due to Peierls for this decrease in conductivity at very low temperatures. He pointed out that the original theoretical expressions had been derived for infinite periodic crystal lattices, and suggested that the unexpected decrease might be due to a retarding influence arising from the reflexion of the elastic waves by the walls of a crystal of finite dimensions. There appears to be a considerable amount of support for this suggestion. Casimir[95] and Makinson[96], using very different methods, have shown that such scattering of the elastic waves at the boundaries of a crystal should result

in a conductivity which becomes proportional to T^3 at very low temperatures. De Haas and Biermasz's experimental value for this index does not yet reach as high a value as 3, but it is certainly tending in this direction and may well do so at still lower temperatures. In addition, subsequent experiments by these workers[97], on crystals of KCl and of diamond, have confirmed that the value obtained for the thermal conductivity does depend on the size of the specimen. At temperatures below the maximum the curves for the various samples diverge, lower conductivity values being obtained as the cross-sectional area of the specimen is reduced. The temperature at which this decrease commences is also higher for the thinner rods.

The value of 3·4 cal./cm.-sec.-°C. now obtained for the conductivity of diamond at −190° C. is no less than seven times greater than that of an earlier determination of Eucken[98]. This worker found very little change in conductivity with increase in temperature, which leads one to enquire whether at normal temperatures this form of carbon is not the best heat conductor available. This may well be the case, as a value as great as 0·85 has already been reported[99] for graphite, another crystalline form of carbon.

Heat conduction in electrical conductors. Some account of Drude's theory of metallic conduction was given in the first volume of these *Reports*[100], when Sommerfeld's application of wave mechanics and the Fermi-Dirac statistics to the problem was outlined. Subsequently, accounts of these theoretical treatments for both thermal and electrical conductivity have been set out in books by Mott and Jones[101] and by Wilson[102], writers who have themselves been active in making further contributions towards the solution of the problem. Drude had shown that the Lorenz function, L (the ratio of the thermal conductivity K to the product of the electrical conductivity and the absolute temperature T), has a value of $3k^2/e^2$, where k is Boltzmann's constant and e the electronic charge. Sommerfeld's treatment gave $L=(\pi^2/3)\,k^2/e^2$ although this result only holds for $T>\Theta$, the characteristic Debye temperature, and is only correct to the first order in kT/ξ and thus for metals such as bismuth and the transition metals for which the thermodynamic potential, ξ, is small, deviations may be expected at high temperatures. The assumption is also made that the whole of the heat conduction takes place by means of electrons, whereas it is now realized that the suggestion originally made by Koenigsberger[103] is true, namely, that heat is also conducted by the lattice. In general, this effect becomes relatively more important for the poorer conductors. Wilson[104] in a later paper explains that the theory of such transport phenomena as heat conduction depends on the solution of the integral equation formed by equating the rate of change in f (i.e. the velocity distribution function of the conduction electrons) due to the external fields and temperature gradients to the rate of change in f due to the retarding mechanism producing the resistance. Denoting this latter rate of change by the collision function $[\partial f/\partial t]_{\text{coll.}}$, it happens with some mechanisms that $[\partial f/\partial t]_{\text{coll.}}=-(f-f_0)/\tau$ where f_0 is the equilibrium distribution and τ the time of relaxation. In cases where this is true the problem becomes comparatively simple, but in general the collision function is an integral operator and it is impossible to define a time of relaxation and a free path. Only at temperatures sufficiently great

for $(\Theta/T)^2$ to be negligible does a free path exist, and in general special methods have to be used to solve the integral equation. Wilson considers the complicated method used by Brillouin[105] to be unsatisfactory, in that it leads to a result disagreeing with the thermodynamic relations between the thermoelectric effects. A more general method involving the use of infinite determinants has been given by Kroll[106]. Wilson himself shows that it is possible to obtain a solution, neglecting only terms of order $(\Theta/T)^3$ by a method of successive approximations. Reference should be made to the original paper for the expressions obtained in this way for the thermal and electrical conductivities, since they are rather too complicated to warrant reproducing in the present account. For an impure metal these expressions lead to Sommerfeld's value for L at very high and very low temperatures and suggest a minimum at intermediate temperatures. Kroll[107] has criticized some of Wilson's conclusions on the ground that they are dependent on the degree of approximation used.

Wilson's method has very recently been carried a stage further by Makinson[96] in a paper which helps considerably in clarifying the position, as direct comparisons are made between the predictions of the theory and experimental results for copper and bismuth. As mentioned in the previous section this author deals also with the conduction of heat by the crystal lattice, a factor which is of little importance for the monovalent metals at any temperature, but for bismuth causes L to assume high values at low temperatures, an effect which is still apparent at normal temperatures.

Experimental results have been published during the past year by Kempf, Smith and Taylor[108] for the thermal and electrical conductivities of a number of aluminium alloys at 20° and 200° C. When the thermal conductivity, K, is plotted against the product of the corresponding electrical conductivity, σ, and absolute temperature, T, it is found that the majority of the points lie close to the line given by $K = 0{\cdot}502 \times 10^{-8} \sigma T + 0{\cdot}03$. Smith and Palmer[109] had previously shown that their results for a series of copper alloys could be correlated in a similar manner, though the actual numerical constants were different.

These results are of considerable practical importance, for since the measurement of thermal conductivity is much more difficult than that of electrical conductivity, it is frequently expedient to derive an approximate value for the thermal conductivity from the simpler measurement. This can be done to a closer degree of approximation by using the appropriate equation for the group to which the alloy belongs than by means of the Wiedemann-Franz-Lorenz law, according to which $K\sigma^{-1}\,T^{-1} = 0{\cdot}58_8 \times 10^{-8}$. The results, however, should still be accepted with a certain amount of reserve. For instance, with aluminium alloys containing a relatively high percentage of silicon, the use of the above equation would give too low a value for the thermal conductivity. This departure is probably due to the very high value of the Lorenz function for silicon, which, at room temperature, according to Koenigsberger and Weiss[99], is some 30,000 times the above-mentioned normal electronic value. Graphite was at the same time shown to have a value of about 160 times normal. Powell[110] has recently obtained a value for the Lorenz function

of a sample of Acheson graphite at 0° C. which is some 200 times normal. The value decreases rapidly with increase in temperature, and over the temperature range 0° to 800° C. it is found that the thermal and electrical conductivities can be represented by the equation $K\sigma^{-1}T^{0.8}=0.0294$, which differs from the Wiedemann-Franz-Lorenz law, both as regards the constant value of the expression and the power to which the absolute temperature is raised.

It would be interesting to see such theories as those described at the commencement of this section applied to this case. Such treatment might reveal whether the high thermal conductivity is entirely due to the high characteristic temperatures (2280° and 760° K.) possessed by this particular form of carbon, or whether the lattice component is unduly important.

Heat conduction in liquid helium. When the properties of liquid helium are completely understood, the chapter in the history of physics which recounts the experimental investigations into the properties of this fluid and their interpretations will make very interesting reading. In the last *Report* mention was made of the first experiments which showed liquid helium II to have an abnormally high thermal conductivity, and it seems appropriate in concluding the present article to endeavour to give some account of the subsequent experimental and theoretical contributions to this subject.

Allen, Peierls and Uddin[111] first published results which indicated the conduction of heat in liquid helium II to be more complicated than was first thought to be the case. They used a sensitive method involving differential helium vapour pressure thermometers to measure the temperature difference between the ends of a capillary tube containing the liquid, and found that the thermal conductivity, when calculated in the normal manner, was a function of the quantity of heat flowing along the tube. At a temperature of 1·865° K., as the heat flow was reduced from 0·883 to 0·52 W./cm.^{2}, the value of this apparent conductivity increased from 3350 to no less than 52,500 cal./sec.-cm.-°C. For a fixed heat flow the conductivity rose to a maximum at about 1·9° K. and then fell steeply again. It was also observed that higher values were obtained for a tube 0·54 mm. in diameter than for one 1·45 mm. in diameter.

A similar dependence of the heat conductivity on the heat flow has since been reported by Keesom, Miss Keesom and Saris[112], who obtained a maximum value of 810 cal./sec.-cm.-°C. at a temperature between 1·7° and 1·9° K. They also note that at still lower temperatures the conductivity decreases, a tendency to approach a value of about 12 cal./sec.-cm.-°C. being reported. Shire and Allen[113], also remark that the conductivity at 0·3° K. may be very small compared with that at 2·0° K.

A number of suggestions have been made to explain these results. The dependence of the conductivity on the heat flow is thought to be due to the destruction of momentum association with this heat flow, whilst the existence of the very marked maximum value is generally assumed to result from the occurrence of some type of ordering transition within the liquid. Fröhlich[114] has directed attention to the similarity between the sharp peak in the specific heat curve in the neighbourhood

of the lambda point to that accompanying an order-disorder transformation, and, assuming helium to have a diamond lattice, has explained the lambda point in terms of a similar disordering which arises from atoms jumping into the vacant points of the underlying body-centred cubic lattice.

Jones[115] has pointed out that the negative coefficient of expansion of helium II supports an explanation of this type, the conditions being likened to those obtaining in a ferromagnetic metal at the Curie point. On this basis it is shown[116] that the high thermal conductivity might be explained in terms of a flow of disorder. Such a flow would result in a similar sharp peak near the lambda point for thermal conduction as for specific heat. Keesom and Taconis[117] have used x-rays to study liquid helium II at these temperatures and consider it to have a very open structure. This structure differs from the diamond lattice suggested by London[118] and by Fröhlich[114], but is one for which very similar reasoning can be applied to explain the specific heat maximum, and one which, on account of its openings, might allow of a large mean free path and high conductivity if conduction takes place by actual mass transfer. Michels, Bijl and de Boer[119] have put forward a very similar suggestion. They consider the heat conduction to result from a transfer of energy by molecules which move through the fluid without being able to exchange their energy with the intervening substance, and show that a mechanism of this type can be deduced from the wave equation.

References

(82) Griffiths, E. *J. Heat. Vent. Engrs*, **5**, 21 (1937).
(83) Weh, J. A. *Gen. Elect. Rev.* **40**, 138 (1937).
(84) Erk, S., Keller, A. and Poltz, H. *Phys. Z.* **38**, 394 (1937).
(85) Dannatt, C. *Engineering*, **144**, 385 (1937).
(86) Powell, R. W. and Griffiths, E. *Proc. Roy. Soc.* A, **163**, 189 (1937).
(87) Wood, W. A. *Proc. Roy. Soc.* A, **163**, 199 (1937).
(88) Debye, P. *Vorträge über die kinetische Theorie*, p. 43 (Teubner, 1914).
(89) Peierls, R. *Ann. Phys., Lpz.*, **3**, 1055 (1929).
(90) Papapetru, A. *Phys. Z.* **35**, 527 (1934).
(91) Blackman, M. *Phil. Mag.* **19**, 989 (1935).
(92) Eucken, A. and Kuhn, G. *Z. phys. Chem.* **134**, 193 (1928).
(93) de Haas, W. J. and Biermasz, T. *Physica*, **4**, 752 (1937).
(94) de Haas, W. J. and Biermasz, T. *Physica*, **2**, 673 (1935).
(95) Casimir, H. B. G. *Physica*, **5**, 495 (1938).
(96) Makinson, R. E. B. *Proc. Camb. Phil. Soc.* **34**, 474 (1938).
(97) de Haas, W. J. and Biermasz, T. *Physica*, **5**, 47 (1938).
(98) Eucken, A. *Phys. Z.* **12**, 1005 (1911).
(99) Koenigsberger, J. and Weiss, J. *Ann. Phys., Lpz.*, **35**, 26 (1911).
(100) Awbery, J. H. *Rep. Prog. Phys.* **1**, 186 (1934).
(101) Mott, N. F. and Jones, H. *Theory of the Properties of Metals and Alloys* (Oxford, 1936).
(102) Wilson, A. H. *The Theory of Metals* (Camb. Univ. Press, 1936).
(103) Koenigsberger, J. *Phys. Z.* **8**, 237 (1907).
(104) Wilson, A. H. *Proc. Camb. Phil. Soc.* **33**, 371 (1937).
(105) Brillouin, L. *Die Quantenstatistik*, p. 357 (Springer, 1931).
(106) Kroll, W. *Z. Phys.* **80**, 50 (1933); **81**, 425 (1933).
(107) Kroll, W. *Sci. Pap. Inst. Phys. Chem. Res., Tokyo*, no. 756, 194 (1938).

(108) KEMPF, L. W., SMITH, C. S. and TAYLOR, C. S. *Trans. Amer. Inst. Min. (Metal.) Engrs*, **124**, Inst. Metals Division, p. 287 (1937).
(109) SMITH, C. S. and PALMER, E. W. *Tech. Publ. Amer. Inst. Min. Engrs*, no. 648 (1935).
(110) POWELL, R. W. *Proc. Phys. Soc.* **49**, 419 (1937).
(111) ALLEN, J. F., PEIERLS, R. and UDDIN, M. Z. *Nature, Lond.*, **140**, 62 (1937).
See also COCKCROFT, J. D., *Nuovo Cim.* **15**, 35 (1938).
(112) KEESOM, W. H., KEESOM (Miss) A. P. and SARIS, B. F. *Physica*, **5**, 281 (1938).
(113) SHIRE, E. S. and ALLEN, J. F. *Proc. Camb. Phil. Soc.* **34**, 301 (1938).
(114) FRÖHLICH, H. *Physica*, **4**, 639 (1937).
(115) JONES, H. *Proc. Camb. Phil. Soc.* **34**, 253 (1938).
(116) ALLEN, J. F. and JONES, H. *Nature, Lond.*, **141**, 243 (1938).
(117) KEESOM, W. H. and TACONIS, K. W. *Physica*, **5**, 270 (1938).
(118) LONDON, F. *Proc. Roy. Soc.* A, **153**, 576 (1936).
(119) MICHELS, A., BIJL, A. and DE BOER, J. *Physica*, **5**, 121, (1938).

OPTICS

Edited by L. C. MARTIN, D.Sc., A.R.C.S., D.I.C.

I. GENERAL INTRODUCTION

BY THE EDITOR

IN accordance with a previous suggestion it has been attempted in certain of the following articles to give an account of progress in subjects which are related to optics but not well covered in the usual physical or optical publications. With this in view a short article on progress in optical instruments and the related theory is followed by contributions on optical glass by Dr W. M. Hampton and Mr W. N. Wheat, on the theory of photography by Mr E. W. H. Selwyn, and on the optical problems connected with cathode-ray television apparatus by Dr W. D. Wright. Acknowledgements and thanks for facilities given in the preparation of the articles are due to Messrs Chance Bros. and Co., Ltd.; to Messrs Kodak, Ltd.; and also to Messrs Electrical and Musical Industries, Ltd.

It is hoped that these latter contributions will be valuable, not only to those immediately concerned with the problems discussed, but also to young physicists now in training who will be able to gather a good idea of the type of technical problem which is likely to be encountered in industry.

II. OPTICAL INSTRUMENTS AND THE THEORY OF OPTICS

BY L. C. MARTIN, D.Sc., A.R.C.S., D.I.C.

The previous reports in this subject covered developments till the end of 1935. A review of optical instruments (such as microscopes) produced during 1936 and 1937 shows a still increasing tendency for them to become adapted as working tools for specific operations. Thus the increasing standardization of production makes operations such as photo-micrography more of a routine operation for unskilled labour rather than the laborious and highly skilled task of a few years ago. The same tendency is particularly noted in spectroscopic apparatus.

A great deal of attention has been given to the development of photometers and spectro-photometers using physical indicators and recording devices. A separate article could well be devoted to this general topic, but the description[1] of the commercial (General Electric Co. of America) model of the Hardy recording spectro-photometer may be cited as a sample of many papers dealing with special instruments.

There is little indication of any other marked tendency in the design of instruments. The Fincham coincidence optometer[2,3,4] may be mentioned as one of the more interesting instruments which have appeared. The action depends on the observation of a coincidence both in the operations of refraction measurement and the location of the meridians of an astigmatic eye. A somewhat similar optical arrangement has been used in a new auto-collimator manufactured by Messrs Taylor, Taylor and Hobson, Ltd., who also manufacture the Fincham optometer.

Reference was made in the previous report to the phase-contrast methods, proposed by Zernike and developed by Burch, which lend themselves particularly to the detection and measurement of irregularities on reflecting surfaces, as for example on an astronomical mirror. An adaptation of the Michelson interferometer with paired auxiliary objectives was suggested some years ago by W. Linnik for the observation of irregularities, microscopic in area, on reflecting surfaces. The plane mirrors of the ordinary Michelson instrument are each replaced by a microscope objective with a plane mirror at its principal focus; if then the illumination is suitably arranged and the objectives are exactly similar, any irregularities of one mirror in relation to the other will be exhibited by fringes in the interference pattern. W. Kinder has described[5] an instrument recently produced on these lines by Messrs Carl Zeiss, and its application in a number of examples, including the measurement of the thickness of individual blood corpuscles deposited on a plane reflecting surface for this purpose.

Interest in special methods of microscopy and photography using ultra-violet[6] and infra-red radiations still continues, and has led to new attention to the designing of suitable lens systems. For this purpose much more complete information as to the refractive indices of optical glasses in the infra-red is sometimes required, and a convenient refractometer for the infra-red has been described[7] by R. Kingslake and H. G. Conrady with an optical system consisting essentially of a double Wadsworth mirror monochromator. The authors give the results of a set of measurements on various optical glasses from 0·365 to 2·6μ.

It is well to remember that many things usually taken for granted in connexion with optical instruments are sometimes open to question. The unexpected discovery made not long ago by Stiles and Crawford, that rays passing through the periphery of the pupil are less efficient in producing illumination of the retina, clearly leads to the conclusion that the brightness of the image produced by an optical instrument will *not* be proportional to the pupil area for a given magnification, as is usually assumed in the text-books. Martinelli[8] has now confirmed this by experiment, and gives numerical particulars. Further, the whole question of Vision in Optical Instruments was dealt with by Prof. Chas. Fabry[9] in the Thomas Young Oration of 1936, in which he pointed out, amongst other things, that the "text-book" type discussion of the resolving limit of an instrument usually rests on the assumption that the object is of a special character, such as two point sources close together or perhaps a grating of alternately opaque and transparent spaces. If then, the image can be shown on diffraction theory to exhibit certain brightness variations corresponding to the object-spacing, resolution should be attained. However, if the

object itself has but small brightness variation, then the corresponding variations in the image may be far beyond the possibility of visual detection, and there will be an enormous discrepancy between the above theory and actual practice; actual visual resolution may only be obtained with much higher magnification. The experimental studies by Arnulf described in Prof. Fabry's address are worth serious attention.

Another subject which might have been considered as settled and yet has received new attention[10, 11] is that of stereoscopic photo-micrography. The essentials of ortho-stereoscopy in this connexion do not seem to have been clearly stated in times past, but micro-stereographs with correct "relief" are shown to have useful applications.

We turn now to theoretical optics. It is satisfactory to find that the recommendations of the Physical Society's *Report on the Teaching of Geometrical Optics* have been so largely acted upon in recent text-books as reviewed by Prof. Allen in *Nature*. Among papers of practical interest, H. Schulz[12] and J. Flügge[13] have continued the discussion of two-mirror optical systems of the type used in dark-ground illuminators, while on the theoretical side Herzberger[14] has contributed papers which seek to throw new light on the fundamentals of optical theory. In the last report, the efforts by Gullstrand and others to formulate an alternative to the usual discussion of collinearity as an introduction to the theory of optical instruments were mentioned. Herzberger has shown that conjugate distance relations corresponding to the usual Gaussian laws can be applied to certain distances measured along skew rays as follows. Imagine a skew ray traversing an axially symmetrical instrument and a plane containing the axis. The paths of the ray in the object-space and image-space respectively will cut the plane in two points which can be called "point" and "diapoint". The ratio of the perpendicular distances of diapoint and point from the axis is called the diamagnification. As the plane rotates around the axis, the points of intersection travel along the rays, and a little consideration will show that the diamagnification will, in general, assume all values. When this magnification is zero, the plane is parallel to the object-space ray, and the intersection of the image-space ray marks the principal diafocal point. When the diamagnification is unity we find the principal points and so on. Relations of the type

$$\frac{n'}{l'} - \frac{n}{l} = \frac{n'}{f'},$$

connect the conjugate distances l and l' measured from the principal points (note that n and n' are the refractive indices and f' is the focal length of the image space), and there are of course valid equations of the Abbe magnification equation type and of the Newtonian as well as of the general Gaussian type. It is not particularly simple to step from this general property to the conjugate distance relations for ordinary paraxial objects and images; Herzberger uses the angle eikonal for this purpose and thus provides an introduction to the theory of the axially symmetrical instrument which avoids at least any misunderstanding with regard to the non-importance of collinear geometry in any fundamental sense. The above work is described as an extension of studies begun in collaboration with H. Boegehold[15]

who, as well as Herzberger[16], has also contributed papers on the problems concerning the physically possible performances of optical systems.

On the measurement of aberrations as apart from their calculation, two useful papers[17, 18] by Kingslake may be cited.

REFERENCES

(1) Michaelson, J. L. and Liebhafsky, N. A. *Gen. Elect. Rev.* **39**, 445–450 (1936).
(2) Biggs, A. H. *Proc. Phys. Soc.* **49**, 444–454 (1937).
(3) Fincham, E. F. *Proc. Phys. Soc.* **49**, 456–468 (1937).
(4) Reason, R. E. *Proc. Phys. Soc.* **49**, 469–478 (1937).
(5) Kinder, W. *Zeiss Nachr.*, 2. Folge, **3**, 91–103 (1937).
(6) Smiles, J. and Wrighton, H. *Proc. Roy. Soc.* A, **158**, 671–681 (1937).
(7) Kingslake, R. and Conrady, H. G. *J. Opt. Soc. Amer.* **27**, 257–263 (1937).
(8) Martinelli, L. *R. C. Accad. Lincei*, **25**, 185–188 (1937).
(9) Fabry, C. *Proc. Phys. Soc.* **48**, 747–762 (1936).
(10) Kurtz, H. F. *J. Opt. Soc. Amer.* **27**, 323–339 (1937).
(11) Martin, L. C. and Wilkins, T. R. *J. Opt. Soc. Amer.* **27**, 340–349 (1937).
(12) Schulz, H. *Z. InstrumKde*, **57**, 251–254 (1937).
(13) Flügge, J. *Z. InstrumKde*, **57**, 495–500 (1937).
(14) Herzberger, M. *J. Opt. Soc. Amer.* **26**, 197–204 (1936).
(15) Boegehold, H. *Z. InstrumKde*, **56**, 98–109 (1936).
(16) Herzberger, M. *J. Opt. Soc. Amer.* **27**, 202–206 (1937).
(17) Kingslake, R. *J. Opt. Soc. Amer.* **26**, 251–256 (1936).
(18) Kingslake, R. *Proc. Phys. Soc.* **49**, 376–380 (1937).

III. RECENT DEVELOPMENTS IN OPTICAL GLASS MANUFACTURE

By W. M. HAMPTON, Ph.D., B.Sc., F.Inst.P. and W. N. WHEAT, B.Sc.
Chance Brothers and Co., Limited

§1. INTRODUCTION

The art of making glass for optical purposes has been treated as a distinct problem for little more than a century. The work of the inventor of the stirring process, Guinand (1748–1824), can be described as of a practical nature only, and the first attempt at a scientific study was commenced by Faraday in 1828[1]. Faraday's discoveries and observations, while of historical interest, lacked real significance, and it was not until 1881, when Schott in collaboration with Abbe commenced his classical investigation, that real progress was made. An account of this work has been given by Hovestadt[2].

The exigencies of the great war undoubtedly stimulated the treatment of optical glass making as a scientific problem and in the post-war period investigators in numerous countries have added to our knowledge of the subject. As regards the position in this country in post-war years reference may be made to papers by Chance and Hampton[3] and Hampton and Wheat[22]. An account of the earlier work of Guinand and Fraunhofer has also lately been given by W. H. S. Chance[4].

The requirements of good optical glass may be stated very simply. It should be a medium which is capable of being optically worked by grinding and polishing to an accurate surface. It should be perfectly isotropic, i.e. it should exert no influence on the light it transmits beyond the characteristic effect exerted by its surface. A wide range of varieties of different refractive powers is required, all having a reasonable degree of resistance to the filming effects of a moist atmosphere. It is to the improvement of these essential requirements that modern researches are directed. The remainder of this article will, it is hoped, give an indication of the trend of a considerable amount of such current research work.

§2. IMPROVEMENTS IN MELTING POTS

The availability of suitable refractory materials for the pots used for melting optical glasses of extreme composition sometimes influences the optical properties obtainable. For example, the new glasses of refractive index 1·62 and reciprocal dispersive power over 60 contain high concentrations of baryta and boric oxide. These high concentrations are difficult to obtain if the molten batch is diluted with excessive quantities of alumina derived from dissolved refractory material. Glasses of this extreme type are notably corrosive and may dissolve 5 per cent or more of their weight of refractory if melted in pots of ordinary fire-clay. Moreover, as the purest pot clays contain at least 0·5 per cent of iron oxide, excessive pot attack inevitably means a high concentration of iron in the resulting glass with consequent increase of light absorption.

The production of super refractory crucibles of laboratory size in the fused or sintered oxides of thorium, magnesium, aluminium, beryllium and zirconium has been developed[5,6,7,8,9], but no information is available that these materials have yet given successful results when used in the manufacture of full size optical glass pots. On the contrary, R. Schwarz and E. Reidt[10] report that the addition of zirconia to kaolin mixtures results in a decrease of refractoriness.

For optical glass in general, and barium crowns in particular, the most successful material so far used for the manufacture of melting pots is sillimanite. Sillimanite, in the pure form regarded as $Al_2O_3.SiO_2$, occurs naturally, and after burning, crushing and grading is mixed with suitable amounts of plastic ball clay for pot making by hand. The percentage of iron oxide impurity can be reduced by acid-washing the crushed mineral.

Sillimanite pot mixtures are more aluminous than the normal fire-clays, and furthermore, when heated to high temperatures decompose to mullite, $3Al_2O_3.2SiO_2$, and silica[11]. Mixtures of alumina and silica do not combine to form mullite below 1600° C., but according to Partridge and Biggs[11] the reaction commences at lower temperatures in the presence of a flux such as molten glass, proceeding slowly with time. The network of mullite crystals formed at the pot-glass interface is highly resistant to glass and appears, when cold, as a white layer. This growth of mullite crystals in a refractory in contact with a lead glass for several weeks has been demonstrated by Partridge and Biggs[12] by means of x-ray analysis.

An alternative to natural sillimanite as an aluminous raw material for optical glass pots is highly burned kaolin or china clay. Using this material greater freedom from iron can be obtained.

The bonding of mixtures rich in sillimanite presents some difficulty if excessive quantities of non-refractory ball clay are to be avoided. The addition of small percentages of felspar has enabled greater mechanical strength and lower porosity to be obtained after burning.

§3. IMPROVEMENTS IN TRANSPARENCY

The chief impurity influencing the absorption of light by colourless glasses is iron oxide. The intensity of the colour produced by a given amount of iron depends on its state of oxidation and also on the type of glass in which it is dissolved. Traces of iron sufficient to produce only a negligible effect in ordinary crown glasses give marked yellow or yellow-green colours in the heavy flint types, and also apparently in the titania glasses, although in the latter case the colour due to titania itself is also influenced by the degree of oxidation[13].

The removal of iron from optical glasses, at least in part, is of prime importance especially when they are intended for use in instruments of long light path. Iron is present as an unavoidable impurity in the batch materials, notably sand, and as stated above is introduced into the glass by solution of the melting pot. In recent years much work has been devoted to the problem of purifying clays and sands. If iron oxide is first reduced to the metal the latter can volatilize as carbonyl or at higher temperatures as chloride. A method of removing iron from sand by means of the simultaneous or consecutive action of sulphur dioxide and chlorine has been patented by Chance Brothers and Co. Limited[14] and a patent[15] also exists for the removal of reduced iron as carbonyl at temperatures below 200° C. Other methods have been devised whereby the clay grog, or sand, is digested in aqueous suspension with hydrochloric acid; with sulphur dioxide[16]; and with acid sodium tartrate solution in the presence of ferrous sulphate[17].

No method has yet been discovered for the complete removal of iron from sands or clays, and by all the methods referred to the amount of iron removed seems to be a function of its original concentration.

A method of removing iron from melting pots by means of chlorine at elevated temperatures has been used with some success[18], although there is some evidence that an increase in the amount of pot attack in the subsequent melting process results. A patent exists[19] claiming the use of chlorine for eliminating iron from molten glass, but the success of this process is open to doubt.

Various additions to the glass batch for the purpose of removing iron during melting have been tried, such as sodium chloride[20] and also a mixture of magnesium chloride and barium peroxide[21].

Although optical glasses are usually melted under oxidizing conditions an equilibrium seems to exist between the proportions of ferric and ferrous ions present. The present authors[22] consider that if the absorption of visible light is to be kept at a

minimum this melting condition is desirable although the vigorous oxidation that seems to be induced in the finished glass by irradiation with ultra-violet or X-rays[23, 24, 25] causes an increase in absorption not only of the ultra-violet but also of visible light[22].

The absorption, at different wavelengths, of visible light by glasses melted under normal conditions has been shown[22] to be very nearly a linear function of the percentage of iron oxide present, and the use of better pots and purer raw materials has resulted in the recent production of glasses of a high degree of transparency.

The following table gives the percentage of light of different wavelengths absorbed by optical glasses from various sources examined recently:

Glass type	Refractive index	Fe per million parts of glass	Percentage absorption per in of glass				
			3650	4358	4916	5460	5790
Borosilicate	1·51	40	4·2	1·4	0·9	0·6	0·6
Barium crown	1·61	55	8·2	4·9	2·1	1·7	2·0
Light flint	1·58	35	13·6	2·1	1·2	0·5	0·6
Dense flint	1·62	12	16·6	1·3	—	0·7	0·8
Extra dense flint	1·74	40	100·0	11·6	3·0	1·2	1·2

§4. NEW TYPES OF OPTICAL GLASSES

Researches on the effect of new or rare glass making materials in recent years have been directed both towards producing glasses of extreme refractive properties and towards improving the durability or surface stability of existing extreme types. An outstanding discovery has been that of the effect of titania. If the composition of a glass be expressed as percentages of the various acidic and basic constituent oxides, the contribution to refractive index of any one oxide is roughly additive over wide ranges of composition. It has recently been shown by Gilard and Dubrul[26] that the refractive index factor for titania is 0·02000 per 1 per cent, a figure appreciably greater than that for any other material so far investigated. The effect of titania on the dispersive ratios of phosphate glasses had been investigated by Harcourt[27], and Stokes and Hopkinson[28]. More recently a patent[29] has been granted covering the use of titania in producing low dispersion glasses with refractive indices in excess of 1·6. It is also claimed[30] that the addition of titania to silicate glasses of high refractive index, such as the heavy flints, enhances their resistance to staining by weak acids.

Clear glasses containing low percentages of zirconia can be made, and it has been stated by Turner and others[31] that this oxide is probably one of the most valuable known for imparting resistance to atmospheric attack.

The literature contains many references to non-silicate glasses. Dennis and Laubengayer[32] describe a high refractive index glass containing germania, and Tillyer[33] mentions a zinc borate glass of low dispersion and refractive index 1·63–1·65. Beryllium glasses, borates[34] and silicates[35], have been prepared with transmissions in the ultra-violet down to 1900 A. Beryllia as a constituent of

ordinary glass in place of lime gives a lower refractive index[36]. It is claimed that glasses of refractive index 1·71 to 2·02 and reciprocal dispersive powers 54 to 19 have been made from the following raw materials, some obviously too rare to be of practical significance for the moment—tungsten oxide, zirconia, lanthana, thoria, and the oxides of tantalum, yttrium, cerbium and hafnium[37].

§5. THE ANNEALING OF OPTICAL GLASS

The problem of attaining a high degree of optical homogeneity in glass by heat treatment, or annealing, has received much attention in recent years. As a result theories on the basic physico-chemical changes occurring at temperatures near the so-called softening point have been profoundly modified.

The fact that a glass which has been cooled rapidly from the plastic state has a refractive index and density below normal has long been recognized. This fact cannot be explained by the mechanical theory of the internal stresses set up in the glass by chilling[38]. It has been stated by F. Eckert[39] that the length of time at which a glass is held near the plastic state influences its refractive index, and Stojaroff[40] has shown that an "equilibrium index" exists for each temperature and when this index is plotted against the temperature the points fall on a straight line. At each temperature the refractive index rises or falls to the equilibrium value according as the initial index is lower or higher than this value. This change of refractive index at each temperature takes place progressively more slowly as the temperature is lowered, and the equilibrium index is higher for lower temperatures of treatment.

A departure from the straight line relationship of equilibrium index and temperature occurs near the so-called transformation point of the glass (the point at which a rapid alteration in expansion coefficient occurs), and at lower temperatures a straight line relationship is again set up roughly parallel to the first but higher by several units in the third decimal place of index[41].

H. R. Lillie[42] has shown that for any temperature corresponding to the "plastic region" of the glass the viscosity also alters slowly to an equilibrium value, increasing, if the glass has first been chilled from the molten state. This he explains by assuming corresponding changes in the state of molecular aggregation, the tendency to association increasing with falling temperature.

The attainment of a uniform refractive index throughout a block of optical glass thus seems to depend on the absence of temperature gradients throughout its mass, except those unavoidably resulting from slow cooling from the temperature selected for the removal of internal stresses[43]. The degree of uniformity required may be gauged from the statement by Finn[44] that a decrease of 1° C. in the effective annealing temperature of a barium crown glass causes an increase in refractive index of 0·000047.

Modern methods of annealing are now based on these discoveries, and it is believed that heterogeneities resulting from the previous heat history of the glass, however severe, can now be removed to the high degree of perfection demanded by the optical instrument maker.

REFERENCES

(1) FARADAY. The Bakerian Lecture 1829; *Philos. Trans.* (1830).
(2) HOVESTADT. *Jena Glass*, translated by J. D. and A. EVERETT (Macmillan, 1902).
(3) CHANCE and HAMPTON. *Proc. Opt. Convention* (1926).
(4) CHANCE, W. H. S. *Proc. Phys. Soc.* **49**, 433 (1937).
(5) *Bur. Stand. J. Res.* **6**, 1131 (1931).
(6) *Trans. Ceram. Soc.* **33**, 33–5 (1934).
(7) *Industr. Engng Chem.* (Anal. ed.), **28**, 486 (1936).
(8) *J. Amer. Ceram. Soc.* **18**, 65–9 (1935).
(9) *Pottery and Glass Record*, **17**, 319 (1935).
(10) SCHWARZ, R. and REIDT, E. *Z. anorg. Chem.* **182**, 1 (1929).
(11) PARTRIDGE and BIGGS. *J. Soc. Glass Tech.* **14**, 63 (1930).
(12) PARTRIDGE and BIGGS. *J. Soc. Glass Tech.* **13**, 130 (1929).
(13) *J. Soc. Glass Tech.* **8**, 187 (1924).
(14) Brit. Pat. 325,386 (1929).
(15) Brit. Pat. 341,060 (1929).
(16) *Sprechsaal*, **70**, 2 (1937).
(17) *J. Glass Soc. Tech.* **19**, 118 (1935).
(18) U.S. Ordnance Dept. Document 2037 (1924).
(19) Ger. Patent, 523,054 (1928).
(20) Russ. Patent, 44,318 (1935).
(21) *Sprechsaal*, **66**, 215–18 (1933).
(22) HAMPTON and WHEAT. *J. Soc. Glass Tech.* **15**, 306 (1931).
(23) *J. Soc. Glass Tech.* **12**, 306, 324 (1928).
(24) *Sprechsaal*, **62**, 314 (1929).
(25) *Trans. R. Soc. Can.* **22**, 433 (1928).
(26) GILARD and DUBRUL. *J. Soc. Glass Tech.* **21**, 476 (1937).
(27) HARCOURT. *Rep. Brit. Ass.* 1871.
(28) STOKES and HOPKINSON. *Rep. Brit. Ass.* 1875.
(29) Ger. Pat. 596,513.
(30) Brit. Pat. 408,993 (1934).
(31) TURNER and others. *J. Soc. Glass Tech.* **11**, 52 (1927).
(32) DENNIS and LAUBENGAYER. *J. Amer. Chem. Soc.* **47**, 1945 (1925).
(33) TILLYER. *J. Opt. Soc. Amer.* **28**, 3 (1938).
(34) Ger. Pat. 585,816 (1933).
(35) *Glashütte*, **62**, 897 (1932).
(36) *J. Amer. Ceram. Soc.* **11**, 535 (1928).
(37) Brit. Pat. 462,307 (1937).
(38) *J. Franklin Inst.* **190**, 597, 835 (1920).
(39) ECKERT, F. *Z. tech. Phys.* **7**, 282 (1926).
(40) STOJAROFF. *Trans. Opt. Inst. Leningrad*, **4**, part 39 (1928).
(41) *Rev. Opt.* (*théor. instrum.*), **5**, 1 (1926).
(42) LILLIE, H. R. *J. Amer. Ceram. Soc.* **16**, 619–31 (1933).
(43) *Z. Instr.* **50**, 155 (1930).
(44) FINN. *J. Opt. Soc. Amer.* **28**, 13 (1938).

IV. ADVANCES IN THE THEORY OF PHOTOGRAPHY

By E. W. H. SELWYN, A.R.C.S., B.Sc., D.I.C., F.Inst.P.
Kodak Research Laboratories

§ 1. INTRODUCTION

The phenomena involved in the production of sensitive material and the final photographic record, and also the properties of the final record are complicated, so that the unravelling of them is slow. Lately a number of apparently unrelated phenomena have been co-ordinated, and in this report they will be dealt with from that point of view. As the interest of the majority of readers will be in the use of photographic materials, the practical implications of recent work will also be mentioned.

There are two general books[1] on photography in English which date back about ten years, and two, rather older[2], which will probably be of greater interest to physicists. Much more voluminous works are available in German[3]. Particular mention may be made of Hay's *Handbuch*, **5** and **6**. The work dealt with here, therefore, is selected from that published within the last five or ten years. The selection has been made with the purpose of showing the correlation of different phenomena, and is probably to some extent a result of personal predilections and prejudices. Much valuable work has been neglected, but this is almost unavoidable.

§ 2. PROPERTIES OF SENSITIVE MATERIALS

Although an attempt has been described[4] to detect the latent image, present in the material before development, it is still true to say that the latent image must be developed before its existence is certain. Thus, although we may frequently refer to the properties of the latent image, there is no certainty that some of the properties in question are not also dependent upon the process of development. The capacity of photographic materials to record small amounts of light is, in fact, dependent upon their developability[5]. A few quanta are able to set up a condition in an individual grain by which the relatively large mass of silver halide composing the grain becomes reducible to metallic silver on development. Recent experimental work[6] suggests that the distribution of the latent image in the grain is not always the same. It is possible by chemical treatment to render the grain undevelopable in ordinary developers. If then the surface of the grain is dissolved away, the remainder becomes developable. It appears, therefore, that the latent image is distributed throughout the grain, but whether the concentration of latent image is uniform throughout the grain or not depends upon the conditions of exposure. Ordinary developers apparently act principally upon the surface of the grain and reduction of the grain to silver depends upon whether there is latent image on,

or very close to, the surface, while physical* developers contain silver bromide solvents and presumably act upon substantially the whole of the latent image. It is possible in fact to dissolve out the silver bromide, and then to obtain an image by physical development, provided that rather more than normal exposure is given in the first place. Thus differences may well be expected in the behaviour of materials according to the developer employed.

When a photographic material has been given a series of increasing exposures and then developed, the density† first rises slowly, then more rapidly and finally reaches a maximum value for ordinary exposures. This effect is shown by a curve, usually known as the characteristic curve, such as in figure 1. From a practical point of view the important features of this curve are its position on the log (exposure) axis, the contrast (or slope of the main part of the curve) and the maximum density. In general, the finer the grain of the material, the farther towards the right does the

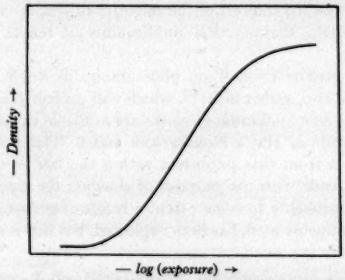


Figure 1. Characteristic curve of photographic material.

curve lie, that is to say the slower it is, and the contrast usually increases as the grain size diminishes. The contrast also depends considerably upon the size distribution of the grains. The shape and position of this curve depends upon the developer used, and also upon the time of development, particularly as regards contrast when the time of development is changed. It also depends upon the colour of the light used for exposing it, and on whether the exposures were made by varying the intensity, or by varying the time, of exposure. Having regard to the variation in shape of the characteristic curve from one material to another and with other variables, it is not surprising that universal agreement has not yet been reached on a standard method of measurement of speed for the purpose of taking pictures in the ordinary way.

The exposure is calculated as the intensity multiplied by the time (It) but these two are not reciprocal quantities, that is to say, for a given colour of light and given development the characteristic curve is not a function of It only. If the logarithm

* Physical developers are so constituted that they deposit silver on any silver already present, whereas ordinary developers reduce the silver bromide to silver.

† Density is defined as the common logarithm of the ratio of the incident to emergent light.

of the exposure necessary to produce a given density is plotted against the logarithm of the intensity of the light, a curve showing a minimum is obtained, as in figure 2. A considerable amount of experimental data of this type was published between 1923 and 1927[7], the general implication of which is that the reciprocity failure curves, as they are termed for brevity, may vary considerably in shape. These curves are usually symmetrical round about the minimum, but break away sharply upwards at very low intensities. There seem to be no general rules correlating the shape of the reciprocity failure curves with other attributes of the sensitive material. All developers give substantially the same shape curve. The maximum efficiency of an exposure occurs when the exposure to produce a medium density is of the order of $\frac{1}{10}$ to 10 sec., and the speed diminishes as the exposure time increases or decreases above or below this value.

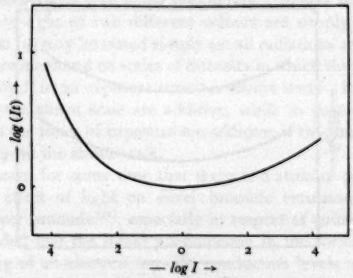


Figure 2. Failure of reciprocity law: the scales are provided to show the order of magnitude of the effect.

One of the most important features of recent investigation of this phenomenon has been the discovery that the reciprocity failure curves, for the same material and time of development, but different colours of monochromatic exposing light, are all parallel to each other[8]. When the values of log (It) necessary to produce a given density are plotted against the logarithm of exposure time, the curves are of the same shape for all wavelengths, but displaced vertically, figure 3. In other words, the shape of the curves depends on the density and time of exposure, and their height on the wavelength and density. Expressed in mathematical form

$$\log It = F_1(D, t) + F_2(D, \lambda), \qquad \ldots\ldots(1)$$

or by rearranging and taking antilogs

$$S(D, \lambda)\, It = \chi(D, t),$$

where $S(D, \lambda)$ may be regarded as the colour sensitivity of the material. Defined in this way the colour sensitivity is independent of the time of exposure, in the sense that the ratio of the amounts of energy necessary to produce a given density, with lights of two different colours, is independent of the time of exposure provided that

the time of exposure is the same for both colours. This result may be of practical use, in photographic comparison of the intensity of spectrum lines, for it means that provided the other conditions implied in the above expression are fulfilled, the calibration need not be made at the same time of exposure as the measurement.

This result also ties up with the discovery that the photographic effect of a mixed radiation composed of two or more monochromatic radiations is related to the photographic effects of the monochromatic radiations by a law[9] which may be written as

$$\Sigma S(D, \lambda_n)\, i_{\lambda_n} t = \chi(D, t), \qquad \text{.....(2)}$$

in conformity with the preceding relation, i_{λ_n} being the intensity of the component of wavelength λ_n in the mixed radiation. Here again $S(D, \lambda_n)$ is inversely proportional to the amount of energy of wavelength λ_n necessary to produce a density D

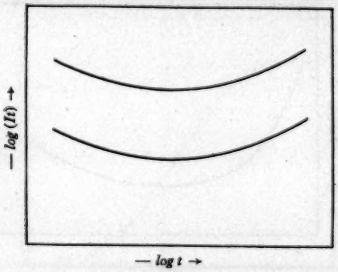


Figure 3. Failure of reciprocity law: curves for two different wavelengths.

at an exposure time t. It is interesting to note that this law had been used, in effect and in a different form, for a considerable time for calculating photographic reflection factors of coloured materials and photographic transmissions of filters, but it was not until 1933 that any explicit experimental proof of its accuracy was made.

In experiments designed in the first place to test the above law of addition, it was also found[10] that, for a given time of exposure t, the ratio of the slopes of the characteristic curves was independent of the density at which the slopes were measured, provided that they were both measured at the same density. That is

$$(\partial \log It/\partial D)_{\lambda_1}/(\partial \log It/\partial D)_{\lambda_2} \text{ is independent of } D. \qquad \text{.....(3)}$$

According to (1) $\log It = F_1(D, t) + F_2(D, \lambda)$,

so that by the relation (3)

$$\frac{F_1'(D, t) + F_2'(D, \lambda_1)}{F_1'(D, t) + F_2'(D, \lambda_2)} \text{ is independent of } D,$$

which requires that $F_1'(D, \lambda) = \psi(D) f(\lambda)$

and $F_2'(D, t) = \psi(D)\, \phi(t)$,

or on integrating and rearranging

$$D = F\left[\frac{\log It - \{f_1(\lambda) + \phi_1(t)\}}{\{f(\lambda) + \phi(t)\}}\right]. \qquad \ldots\ldots(4)$$

The remarkable conclusion[11] is therefore reached that the shape of the characteristic curve obtained by varying the intensity, keeping the time of exposure constant, is independent of the time of exposure or colour of the light, except in so far as the curve is stretched along the log-exposure axis, according to the factor $\{f(\lambda) + \phi(t)\}$ and located on the same axis by $\{f_1(\lambda) + \phi_1(t)\}$. The range of validity of the general relation (4) is doubtful, but if it should prove to be wide, then comparatively few measurements will suffice to provide all the sensitometric properties of a material. It has also been found that, provided the intensities of illumination are adjusted to give equal photographic effects in equal times, successive exposures to light of two different colours are simply additive[10]. Both this and the relation (2) may be stated simply as: all radiations are photographically equivalent if they are measured on scales of intensity in which the intensity necessary to produce a density D in an exposure time t is always unity. In mixed radiations, the intensities on the above scale are additive, while in successive exposures to different radiations the times of exposure are additive, if the total time is t and the intensities are unity on the above scale.

It has been known for some time that there is a striking parallelism between the photographic effect of light on silver bromide emulsions and the photoconductivity of silver bromide[12], especially in respect of colour sensitivity, from which it is concluded that the initial phenomenon in the formation of the latent image is the raising of an electron into the conduction levels of the crystal. The above results are in agreement with this view. It may be said, on the strength of them, that the colour sensitivity depends upon the capacity of light of different wavelengths to raise the electrons into the conduction levels. Whatever happens afterwards, and this is determined partly by the rate at which electrons are raised, is independent of the colour of the light used for exposure. In other words, an electron in the conduction levels is an electron in the conduction levels, whether it was blue or red light that put it there. In the above connexion it has been shown that there is a strong resemblance between the absorption spectrum of the silver halide grains and the colour sensitivity of the emulsion[13]. Thus, in discussing the action of light on photographic materials, colour sensitivity may be regarded as a phenomenon separate from others, such as reciprocity failure and the shape of the characteristic curve. Although other means of changing colour sensitivity are known, sensitivity to colours other than blue is always conferred in practice by adding dyes to the emulsion. Their mode of action appears still to be uncertain, but a considerable number of dyes with this property have been discovered. For instance, there has been a steady increase in the long wavelength limit of sensitivity of commercially available infra-red sensitive plates since their introduction in 1925[14]. This continual extension cannot continue, in practice, indefinitely, for the existence of thermal energy sets a limit in the neighbourhood of 2000 mμ[15].

A considerable amount of literature has accumulated concerning the intermittency effect, but the explanation of the effects observed now appears to be relatively simple[16]. Imagine the light which is used for exposing the material to be interrupted by a sector, of which the speed is varied, and the amount of energy (It) which is necessary to produce a specified density to be plotted against the number of interruptions per second. Then a curve such as in figure 4 is obtained, and it has been found that the lower region corresponds with the amount of energy necessary to produce the specified density at the intensity level of the uninterrupted light, while the upper region corresponds with the amount of energy necessary to produce the same density with continuous illumination when the intensity is equal to the average from the interrupted light. The intermittency effect is thus a manifestation of reciprocity failure. The curve obtained rises or falls according to the intensity level and the shape of the reciprocity failure curve. The correspondence in the lower part of the curve might have been expected, since it means that there is no significant

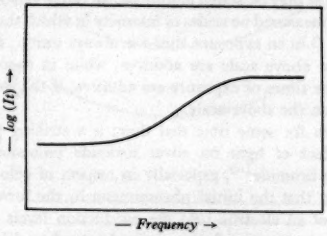


Figure 4. Intermittency effect.

change in the latent image during the relatively short dark intervals. The correspondence in the upper part is attributed to the quantum-like structure of the light, so that if the interruptions are fast enough, there is no essential difference between interrupted and non-interrupted illumination. The interruptions need not be so fast as at first sight might be expected, for the number of quanta falling on a grain in 1 sec. is small, even at high intensities of illumination. It is, in fact, possible to calculate, from the frequency at which the interrupted illumination is equivalent to continuous illumination of the same intensity, the mean target area for quanta, and it turns out that this is of the same order as the area of the single grains of the emulsion. In support of the above explanation detailed calculations have been made[17] of the time-distribution of quanta in interrupted and non-interrupted illumination and it was predicted from them that the frequency with which the light must be interrupted to simulate continuous exposure should vary with the ratio of light to dark periods in a manner which was confirmed by experiment. From a practical point of view, the result may be of value since it means that sectors revolving at a high enough speed are equivalent to neutral filters. Similar arguments may possibly be applicable to revolving sectors used for visual experi-

ments, and with photoelectric cells. So far as colour vision is concerned a similar rule for mixtures of light seems to hold as is described above in equation (2) for their photographic effect.

An outline of the evidence that the first action of light is to raise electrons into the conduction levels of the crystal has already been given. In so much as the subsequent phenomena are very much dependent upon the time of exposure, it may reasonably be surmised that particles heavier than electrons are involved, in which case the shape of the reciprocity failure curves might be dependent upon temperature. This has indeed been found to be the case[18]. As the temperature is lowered, the reciprocity failure curve changes its shape. This change of shape readily accounts for the apparent contradictions in previous observations on the change of sensitivity with temperature. The nature of the change is such as to suggest that two phenomena are involved in producing the curve with a minimum which is observed at ordinary temperatures.

An explanation in very general terms has been attempted for the two separate phenomena and the separation of the reciprocity failure curve into two parts has suggested a formula for the curve which fits the observations extremely well, but the position is not yet quite satisfactory[18, 19]. Experiments which have not yet been published in full have shown that the reciprocity failure curve is flat at liquid air temperatures. Presumably at that temperature the more massive particles, probably silver ions, cannot move, and only move when the exposure is completed and the temperature raised for development. Provided that the total energy input of a fixed colour is the same, the time of exposure is immaterial[20]. Recent theoretical arguments[21], based on the various phenomena associated with electrical conductivity in alkali and silver halides, will probably lead to a much clearer understanding of the processes which take place in the silver halide grain. For some time now the latent image has generally been considered to consist of minute specks of metallic silver, that have coagulated upon "sensitivity nuclei" which may be impurities like metallic silver, or silver sulphide, or discontinuities in the lattice. The evidence for this is mainly chemical, and consideration of it and the theoretical arguments may legitimately be postponed until our knowledge of latent image formation is more complete.

It has already been mentioned that simple additive laws apply to exposures to mixed radiations and successive exposures to different radiations, under certain restrictions, but when these restrictions are not adhered to very curious effects may happen. Although there has been no very considerable advance in our knowledge of them in the last five years, they may be worth while mentioning, especially as they may be met with in the practical use of photography.

It appears that if a first exposure is made by which a considerable number of electrons is suddenly raised into the conduction levels, and a subsequent exposure is made in which electrons are raised more slowly, the density obtained is less than would be expected from the densities obtained when the exposures are made separately. Thus a flash of lightning recorded on a film which is afterwards fogged, as for instance by skylight, may appear white on a dark background on the negative

(Clayden effect) and similar effects are produced by exposure to x-rays and then to white light of moderate intensity (Villard effect). A similar effect is produced by pressure, by writing on the film for example, and subsequent exposure, but in this case, although there may be sudden release of electrons, this has not been proved.

The summation of the exposures is as near normal as can be estimated if the order of exposures is reversed, so that it must be concluded that very brief exposures to light of high intensity, or exposures which simulate these conditions, actually reduce the sensitivity of the material to exposures of longer duration and lower intensity. Whether these effects will be of real importance to understanding the process of formation of the latent image cannot be said, but it has been suggested[22] that they may be related to the loss of sensitivity for very brief exposures, as shown by the rise of the reciprocity failure curve in that region. At the moment it is supposed that these effects are attributable to differences in distribution of the latent image for short and long exposures[23].

§ 3. PROPERTIES OF THE DEVELOPED IMAGE

The density of a developed photographic material depends upon the manner in which it is measured, owing to the fact that the developed silver grains scatter the light. Thus if a collimated beam of light is caused to fall on the material and all the emergent light is collected, the density so measured (diffuse density) is less than if only the light which still remains collimated after emergence is collected (specular density). Detailed calculations and experiments[24] were made on this point some years ago, and the main outlines of the phenomenon can be regarded as clear. In general the ratio of specular to diffuse density increases as the size of the grains increases. This has been made the basis of a method of measuring the average grain size of developed materials, with the object of providing a relatively simple method of estimating the graininess of negative materials[25]. The number of grains per unit area and the amount of silver per unit area may also be determined by means of this method. While the simplicity of the measurement is attractive, there is still some doubt whether there is really close correlation between the results obtained and results obtained by other more direct measurements of graininess.

Until comparatively recently the graininess of photographic materials was measured by visual processes, the nature of which was not completely understood. With the advent of microphotometers for the measurement of spectrum lines and similar fine detail in photographs a new point of view has been commonly adopted. Imagine a photographic deposit of uniform density divided up into numerous small equal areas. The graininess may then be considered as completely described if the statistical distribution of density among these small areas is known for all sizes of the small areas. Several attempts[26] have been made to construct a theoretical description of the nature of this variation, but none of them is really satisfactory. It is not very difficult to calculate with a fair degree of approximation to the truth the distribution of grains which would be expected on the supposition that they are distributed at random and also on the supposition that the developer does not

interfere with this randomness of distribution. The transformation of this distribution into one of density distribution has not been achieved satisfactorily. The point of difficulty is that although it is reasonable to take the density as proportional to the concentration of grains when considering large areas, for small areas the space distribution of the particles must also enter into the calculation of density. In view of this difficulty a different mode of attack has been attempted in which it is supposed that the graininess is dependent upon one specific constant of the material[27]. The argument, which is of interest in view of the fact that density in the ordinary way would be considered to have zero dimensions, is based upon the theory of dimensions as developed by Bridgman[28]. Whichever method of calculation is adopted it turns out that the variation of density should be at any rate roughly proportional to the grain size, and inversely proportional to the square root of the area over which each density measurement is made. Numerous measurements have been made of this statistical distribution of density[29]. There is some variation between the results but they can all be brought into line if it is supposed that the actual distribution does not depend exactly on the square root of the area of the scanning spot. What this means is still obscure and the relation between visual estimates of graininess and measurements of the distribution of density has only been suggested theoretically.

In connexion with the problem of the influence of graininess on microphotometer records it may be noted that owing to the combined effects of the rise in graininess with density and measurement of the microphotometer of the transmission the maximum variation in the records due to graininess occurs where the transmission is about 0·5. This, no doubt, is just about where it is least desirable.

An attempt has been made to calculate the influence of the lag of the galvanometer on the irregularity introduced by graininess into a microphotometer record[30], but it is doubtful whether the calculation does in fact give the same results as are obtained in practice, for it involves the assumption that the correlation between small areas is greater, the greater their distance apart, whereas exactly the opposite might be expected. In any case, the lack of resolution introduced by the galvanometer or other recording mechanism in a microphotometer must be kept small compared with the detail which it is desired to measure, and the variations due to graininess must be accepted as inevitable. Such variations can usually be recognized if twin photographs are taken[31].

A considerable number of developers has been devised with the object of producing less graininess than would be obtained with more normal developers. Some of these developers are of little value, and in any case it appears certain that when a developer does give decreased graininess the speed of the material and its contrast diminish. Accordingly it would appear to be more satisfactory, in general, to use a finer-grained material in the first place even if it is slower, since processing with finer-grain developers is usually more difficult than with standard developers.

So far as other properties of the developed image are concerned, there has been little further progress, except in the purely practical advantages associated with improvements in sensitive materials. From the user's point of view it may be

worth mentioning that materials made especially for scientific purposes, such as spectroscopy, astronomy, oscillograph recording and radiology are available. In respect of the first two uses a recent series of papers[32] contains much useful practical information.

REFERENCES

(1) Neblette, C. B. *Photography, its Principles and Practice* (1927, 1931).
Clerc, L. P. (edited, Brown, G. E.). *Photography, Theory and Practice* (1930, 1937).
(2) *Photography as a Scientific Implement* (1924).
Ross, F. E. *Physics of the Developed Photographic Image* (1924).
(3) Eder, J. M. *Handbuch der Photographie* (1912–32).
Hay, A. *Handbuch der Photographie* (1929–33).
(4) van Kreveld, A. and Jurriens, H. J. *Physica*, **4**, 285 (1937).
(5) Eggert, J. *Photogr. J.* **76**, 17 (1936).
(6) Meidinger, W. *Phys. Z.* **36**, 312 (1935); **38**, 564 (1937); **38**, 737 (1937).
Kempf, A. *Z. wiss. Photogr.* **26**, 235 (1937).
(7) Jones, L. A. and Webb, J. H. *J. Soc. Mot. Pict. Engrs*, **23**, 142 (1934). (This contains a general account of experimental results up to 1934, and references to earlier publications.)
Arens, H. and Eggert, J. *Z. Phys. Chem.* **131**, 297 (1928).
(8) Webb, J. H. *J. Opt. Soc. Amer.* **23**, 316 (1933).
(9) van Kreveld, A. *Physica*, **1**, 60 (1933); *Z. wiss. Photogr.* **23**, 222 (1934).
(10) Webb, J. H. *J. Opt. Soc. Amer.* **26**, 12 (1936).
(11) van Kreveld, A. and Ornstein, L. S. *Proc. Acad. Sci. Amst.* **39**, 477 (1936).
(12) Toy, F. C. *Proc. VII Int. Cong. Photogr.* p. 23 (1928).
Toy, F. C. and Harrison, G. B. *Proc. Roy. Soc.* A, **127**, 613 (1930).
(13) Leermakers, J. A., Carroll, B. H. and Staud, C. J. *J. Chem. Phys.* **5**, 878 (1937).
Leermakers, J. A. *J. Chem. Phys.* **5**, 889 (1937).
(14) Brooker, L. G. S. and Keyes, G. H. *J. Franklin Inst.* **219**, 255 (1935).
Dieterle, W. and Riester, O. *Z. wiss. Photogr.* **26**, 68, 141 (1937).
See also Ref. (32).
(15) Czerny, M. *Z. Phys.* **53**, 1 (1929).
(16) Webb, J. H. *J. Opt. Soc. Amer.* **23**, 157 (1933).
(17) Silberstein, L. and Webb, J. H. *Phil. Mag.* **18** (ser. 7), 1 (1934).
(18) Webb, J. H. *J. Opt. Soc. Amer.* **25**, 4 (1935).
(19) de Langhe, J. E. *Physica*, **3**, 904 (1936).
(20) Berg, F. W. and Mendlessohn, K. *Proc. Roy. Soc.* A, **168**, 168 (1938).
(21) Webb, J. H. *J. Opt. Soc. Amer.* **26**, 378 (1936).
Mott, N. F. *Sci. J. R. Coll. Sci.* **7**, 134 (1937).
Mott, N. F. and Gurney, R. W. *Proc. Roy. Soc.* A, **164**, 151 (1938).
Sheppard, S. E. *Photogr. J.* **68**, 413 (1928).
(22) Luppo-Cramer. *Photogr. Korr.* **72**, 1 (1936).
(23) Landau, R. *Sci. Ind. Photogr.* **2** (2nd ser.), 1 (1931); **5** (2nd ser.), 137 (1934).
(24) Tuttle, C. and Silberstein, L. *J. Opt. Soc. Amer.* **14**, 365 (1927).
(25) Eggert, J. and Küster, A. *J. Soc. Mot. Pict. Engrs*, **30**, 181 (1938). (This gives a general account of all previous work and includes references.)
(26) Bricout, P. *C.R. Acad. Sci., Paris*, **197**, 1202 (1933).
Kotowski, P. *Kinotechnik*, **16**, 209 (1934).
Albersheim, W. J. *J. Soc. Mot. Pict. Engrs*, **29**, 417 (1937).
Siedentopf, H. *Phys. Z.* **32**, 454 (1937).
(27) Selwyn, E. W. H. *Photogr. J.* **75**, 571 (1935).
(28) Bridgman, P. W. *Dimensional Analysis* (1931).
(29) Bricout, P. *C.R. Acad. Sci., Paris*, **197**, 1202 (1933).
van Kreveld, A. *IX. Congres. Intern. de Photogr.* (*Paris*), p. 263 (1935); *J. Opt. Soc. Amer.* **26**, 170 (1936).

VAN KREVELD, A. and SCHEFFER, J. C. *J. Opt. Soc. Amer.* **27**, 100 (1937).
SIEDENTOPF, H. *Phys. Z.* **32**, 454 (1937).
GOETZ, A. and GOULD, W. O. *J. Soc. Mot. Pict. Engrs*, **29**, 510 (1937).
(30) VAN KREVELD, A. and ORNSTEIN, L. S. *Physica*, **3**, 815 (1936).
(31) CHALONGE, D. *Reunions de l'Institut d'Optique 4th year*, p. 77 (1934).
GEORGE, W. H. *Proc. Phys. Soc.* **49**, 357 (1937).
(32) MEES, C. E. K. *J. Opt. Soc. Amer.* **21**, 753 (1931); **22**, 204 (1932); **23**, 229 (1933). (The substance of these papers has been reprinted as a booklet by Eastman Kodak Co.)

V. TELEVISION OPTICS

BY W. D. WRIGHT, D.Sc., D.I.C., A.R.C.S.

§ 1. MECHANICAL OPTICAL SYSTEMS

There are two major difficulties inherent in mechanical-optical television systems. These are (1) the problem of getting enough light through the system and (2) the reduction in size of the rotating parts to avoid the use of excessive power to maintain accurate synchronization. In an optical scanner, an increase in definition calls for an increase in the number of holes, lenses or mirrors, as the case may be. If the mechanism cannot be increased in size because of synchronizing problems, then the components must be made smaller. When this is done, the overall light efficiency of the system falls very rapidly and the effect is very pronounced in present-day high definition television. These difficulties are now well understood[1] and need not be enlarged upon here.

Attempts to overcome them have been made in several directions, as in the development of special scanning systems, through the use of very high intensity light sources, in the design of new light modulating devices and so on. Two systems of this type which call for particular mention are the Scophony system and the Mihaly-Traub system. In the Scophony system, use is made of supersonic vibrations transmitted through a liquid cell from an oscillating quartz crystal. The vibrations form, in effect, a diffraction grating in the cell, in which each point in the cell corresponds to a different element of the picture. By modulating the vibrations with the incoming signals, the amount of light diffracted at each point is regulated, and a number of picture points are projected simultaneously on to the screen. This leads to greater brightness of picture and simpler scanning apparatus. For further details of the device with its associated optical system, reference must be made elsewhere[2]. In the Mihaly-Traub system, a small mirror drum rotates inside a hollow stationary drum. The rotating part is very much smaller than in the simple mirror drum system and the light efficiency as a whole is very much greater. The details of the system are described in[3].

§ 2. CATHODE-RAY APPARATUS

The successful establishment of a television service has introduced a number of problems of special interest to the optical designer. A few years ago optical apparatus formed an important part in all television systems, but the development

of electron scanning devices and the cathode ray tube receiver has meant that many of the problems formerly associated with mechanical-optical systems and with light modulating devices have been transferred to the field of electron optics. In this account, the discussion has been limited to the optical questions that arise in the system of transmission now in use by the B.B.C. at the Alexandra Palace and, in the case of receivers, to the sets now on the market capable of receiving the programmes that are being broadcast.

The Emitron camera. Two forms of electron camera are in use, their essential features being illustrated in figures 1 and 2. In the device shown in figure 1, an image of the scene to be transmitted is focused with the lens *L* on the photoelectric mosaic *M* through the window *W*. Each photoelectric element becomes positively charged in proportion to the amount of light falling on it, and the signal which is to be transmitted is obtained by the discharge of successive elements as a narrow

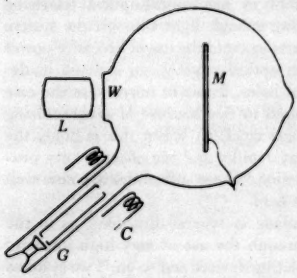


Figure 1.

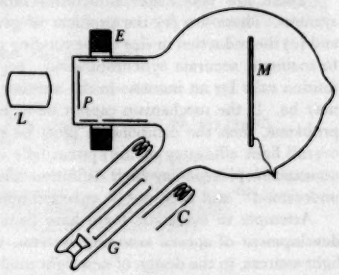


Figure 2.

beam of electrons projected by the electron gun *G* sweeps continuously across the mosaic. The scanning of the electron beam is produced electro-magnetically by the coils *C*.

The main optical problem concerns the lens *L*. This is a normal type of photographic objective capable of covering the field delimited by the size of the mosaic, with a definition determined by the detail which the television system as a whole can resolve. It will be obvious from a comparison with the needs of the cinema, that it would be highly desirable if a turret of lenses could be attached to the camera, to provide the operator with a choice of focal lengths and apertures. Unfortunately, the geometry of the camera and the view finding and focusing problem make this a very difficult demand to meet. The dimensions of the camera, which are governed by electrical factors and manufacturing considerations, are approximately as follows:

Size of mosaic	4 × 5 in.
Distance from mosaic to window ...	5 in.
Diameter of window	2½ in.

Obviously the lens cannot be mounted very far from the window, otherwise the edge of the latter would produce a vignetting effect on the image. The diameter of the window is limited, since the greater its diameter, the farther must the electron gun be mounted off the axis of the system. The obliquity of the gun in any case introduces a keystone distortion on the mosaic which has to be corrected in the scanning circuits, while the impossibility of focusing the electron beam sharply at all points of the oblique mosaic has to be allowed for by using a very narrow pencil of electrons to give a correspondingly greater depth of focus. As the window is made wider, so these factors become more troublesome.

The only way to vary the focal length and at the same time to retain the lens close to the window is to use telephoto lenses. A reversed telephoto lens with its principal planes inside the electron tube would provide a possible way of getting a shorter focus lens, but the increased angular diameter of the field limits the application of the method. A normal telephoto lens, with principal planes in front of the objective, can be employed without difficulty to give a longer focal length, at least up to a value of 12 in. focus.

The definition required presents no special difficulty. The number of scanning lines in the picture is approximately 400 and on the mosaic these lines are spread over a distance of 4 in. Each line is therefore 0·01 in. wide and in order that the lens should contribute no appreciable loss of definition to the final picture, the disc of confusion should not exceed about 0·0025 in.

It is, however, not easy to specify the performance of a lens so far as definition is concerned in a manner that is well suited for television purposes. This applies equally to lenses used at the transmitter and the receiver. The number of lines per inch that could be resolved would be suitable, were it not for the fact that lines can still be resolved even though the contrast between the dark and light spaces is very poor. This loss of contrast is serious and should be known, and if it could be stated quantitatively for various numbers of lines per inch, it could be expressed as an attenuation curve, in decibels if necessary. As frequency attenuation in the electrical circuits of the system in effect produces a loss of contrast in the finely resolved parts of the picture, this method of recording definition would be ideally suited for the purpose in mind. It is also possible that the same method of specification would be valuable in other directions, although it must be admitted that the theoretical and experimental evaluation of contrast in fine detail presents considerable difficulties.

The aperture of the lens cannot usefully be greater than $f/3$. This is governed partly by the diameter of the window, but mainly by depth of focus considerations. Owing to the comparatively long focus objectives that have to be used, the depth of focus in the object space is rather small. Further, the mosaic, which consists of minute, sensitized silver globules deposited on mica, can be kept flat only to within relatively coarse limits, and the use of a lens with an aperture of $f/2$ or $f/1{\cdot}5$ would be liable to produce an image that could be focused accurately over only part of the mosaic at a time. Similarly, the squaring-on of the mosaic and lens would become a very delicate operation. Hence, although high apertures would be

desirable from the point of view of the light sensitivity of the system, they are not feasible in practice.

The chromatic correction of the lens should be adjusted to suit the energy distribution of the lighting used on the scene and the spectral sensitivity of the photoelectric mosaic. If any one camera is likely to be used both for scenes illuminated artificially and for outdoor scenes, then obviously no special design can be adopted to suit one or other, but a compromise is necessary. When artificial illumination alone is employed, then some small gain in definition might be obtained by using lenses designed to allow for the smaller amount of energy at the short wavelengths. The spectral-sensitivity of the mosaic is far more important. The actual sensitivity curve will depend on the nature of the photoelectric surface, and may be either blue sensitive, be somewhat similar to the luminosity curve of the eye, or may possess considerable sensitivity in the infra-red. In each case the appropriate chromatic correction should be applied to the lens, to bring the shortest focus in the secondary spectrum to the same wavelength as the maximum of the sensitivity curve of the mosaic.

One of the main difficulties in the way of the provision of a turret of lenses is connected with the problem of focusing the object on the mosaic and with the use of a viewfinder. It is possible to focus the image fairly accurately by viewing it directly on the mosaic, but this is not a convenient operation and is in any case quite unsuitable for viewfinding purposes. A comparison with cinema technique is not altogether legitimate in this case. It is not unusual in a cinema studio for the distance from the camera to the central figure on the set to be obtained directly with a measure. The camera can then be focused by means of the previously calibrated focusing scale. This method can only be applied when the shot to be taken is of comparatively short duration and the movement in the scene is limited. Most cinema shots are, generally speaking, quite short, but a television programme may continue for an hour without a break. The possibility of measurement is therefore out of the question; the operator has to be constantly following the scene in the viewfinder and constantly checking the accuracy of his focus. For this reason, the focusing and viewfinding system has to be as simple as possible and such as to produce the minimum of fatigue to the operator; at the same time the accuracy of focusing has to be high.

A rangefinder type of focusing, although it gives the necessary accuracy, is not suitable for continual viewing, especially when the viewing has to combine the double function of viewfinding and focusing. The method adopted has been to employ a second lens of the same focal length and aperture as that used to focus the image on the mosaic, in conjunction with a ground glass screen of the same dimensions as the mosaic. The two systems are then mounted side by side, as illustrated in figure 3 *a*. The two lenses can be linked together mechanically and the focusing of one lens can be made to focus the other, provided the lenses are properly matched and that a preliminary setting has been made to ensure that the mosaic and glass screen are simultaneously in focus for one given object distance. One complication arises in this system due to the width of the electron camera. Because

of this the two lenses cannot be brought closer together than some 6 or 8 in., and when focusing near objects, this separation will introduce very appreciable parallax unless some method of correction is employed. The correction can be obtained either by rotating the axis of the viewfinder system and advancing it towards the object to maintain the same distance between object and both lenses, figure 3*b*, or alternatively, to move the viewfinder lens laterally inwards as objects nearer to the camera are focused, figure 3*c*. Both methods make the mechanical linkage system more intricate and while the system in figure 3*b* is perhaps to be preferred optically, that in figure 3*c* is mechanically the simpler.

It will be readily appreciated from this short account how difficult it would be to devise an arrangement incorporating a turret of lenses in front of both camera and viewfinder and to maintain the focusing register as the system was changed from one lens to another. These obstacles have not yet been overcome. Fortunately,

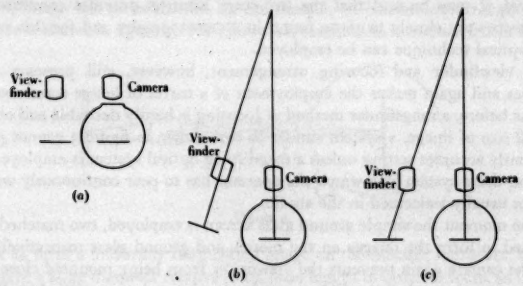


Figure 3.

it is frequently advantageous to employ two or more cameras, working simultaneously at different distances and angles from the scene and, if necessary, with different lenses. The picture finally transmitted can be selected from any of the cameras and can be electrically faded from one to the other. This, in part at least, overcomes the lens turret problem.

The second type of camera, figure 2, gives rather more optical freedom than the first type. The operation of the tube is briefly as follows: An optical image of the scene to be transmitted is focused on the photoelectric surface *P*. An electron optical focusing system *E* forms an electron image of the photocathode on the mosaic *M*. Each point of *P* emits electrons in proportion to the light falling on it and when focused on *M* these electrons liberate a proportionate number of secondary electrons and leave the mosaic elements positively charged as before. The electron beam from the gun *G* sweeps continuously across the mosaic, discharging each element in turn, to give the signal impulses which are transmitted.

This form of tube has a greater light sensitivity than the former, owing to the fact that each primary electron on the mosaic liberates a number of secondaries,

but when ample light is available to work either tube, the first type is at present usually employed.

Optically, the two-stage Emitron has several advantages. In the first place, the photocathode is sufficiently close to the window so that there is no material restriction on the focal length of the camera lens to be used. Again, the size of the cathode, of the order of $1\frac{1}{2}$ in. diameter, is such that the field can be covered satisfactorily with lenses of very short focal lengths, and as a result advantage can be taken of the greater depth of focus in the object space that can be obtained with these lenses. Alternatively, a higher aperture lens can be used, if the potential gain in depth of focus can be sacrificed. Lenses with an aperture of $f/1{\cdot}4$ have, in fact, been employed. There is the further advantage that the angular diameter of the field is no longer limited as in the tube in figure 1, where the large mosaic placed well back from the window inevitably fixes the angle of view within fairly narrow limits. In general, it may be said that the two-stage Emitron provides conditions that approximate very closely to those found in cinematography and for this reason a similar optical technique can be employed.

The viewfinder and focusing arrangement, however, still presents certain difficulties and again makes the employment of a turret of lenses somewhat awkward. As before, a rangefinder method of focusing is hardly desirable and owing to the small size of image, a system similar to that shown in figure 3 cannot provide a sufficiently accurate setting unless a magnifying optical system is employed. For prolonged use a system into which the operator has to peer continuously with one eye is not usually welcomed in the studio.

At the moment the simple ground glass screen is employed, two matched lenses being used to form the images on the mosaic and ground glass respectively. The size of the camera again prevents the viewfinder from being mounted close to the camera lens, so that as before a linkage mechanism has to be introduced to allow for parallax.

§ 3. FILM-TRANSMISSION

In the normal cinematograph projector each frame is projected for a given time, after which a shutter obstructs the light and simultaneously the film is moved to the next frame. To remove the flicker the film is also shuttered while stationary, with the additional effect of reducing the brightness of the final picture. In the Emitron camera each element on the mosaic becomes charged to an extent depending on the total amount of light falling on the element between successive scans of the mosaic by the electron beam. If the normal, shuttered image from a cinema projector were focused on the mosaic, it might be expected that, as with the eye, the only action of the shutter apart from screening the motion of the film would be to reduce the illumination on the mosaic. Unfortunately, the sudden change from a high level of illumination to zero illumination as the shutter rotates produces a temporary disturbance to the potential on the mosaic and hence to the signals transmitted; to avoid this, some modification of the system becomes necessary.

One alternative that has proved fairly successful consists of illuminating the mosaic for only a small fraction of the normal duration of a frame. The mosaic integrates or "memorizes" this illumination until each element is discharged with

the scanning beam, after which the whole process is repeated with the next frame. The outstanding optical problem that remains is the method of imaging each frame of the film, approximately 0·875 in. × 0·75 in. in size, on to the mosaic 4 in. × 5 in. The camera used in the film transmitter is that illustrated in figure 1. We have already seen that the distance from the lens to the mosaic is largely governed by the distance of the window from the mosaic, and since the magnification can be calculated from the figures just quoted, the distance from the film to the lens and hence the distance from the film to the camera, can be found. This distance, therefore, depends on the previously fixed dimensions of the television camera and the cinema film, and can only be varied within narrow limits. It can easily be seen that the distance will be very short, so short in fact that it is quite impossible to bring the electron camera sufficiently near the cinema projector for the system to work. For this reason an intermediate image has to be formed and from this image a second, enlarged, image has to be thrown on the mosaic. A convenient arrangement for doing this is shown in figure 4. It is, however, essential to introduce a field lens, L_1,

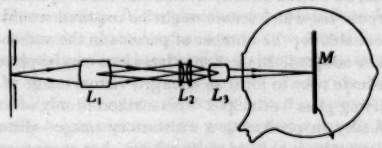


Figure 4.

in order to form a uniformly illuminated image on the mosaic; moreover, this lens has to have a large diameter relative to its focal length in order to focus an image of L_1 in L_3. It can easily be seen that this high aperture is ultimately determined by the size of the mosaic relative to its distance from the window. The field lens, which has to be achromatic, introduces considerable Petzval curvature of field, making it difficult to secure a sharp image over the whole of the mosaic.

It is interesting to note that, like so many other optical problems occurring in industry, the difficulties arise almost entirely from the predetermined geometry of other parts of the apparatus. Regarded purely as an optical problem, the system would have no unusual or troublesome features.

A different method of overcoming the sudden alternations in intensity on the mosaic is to employ a continuous motion film projector in place of the normal intermittent type. The general principle of the continuous motion projector is that a stationary image is projected on the screen from a continuously moving film by means of a rotating lens or mirror system. The illumination of the screen is practically constant since as the light from one frame decreases, the light from the next increases.

A Mechau projector, which employs a large disc of eight mirrors, has been adapted to project an image into the Emitron camera. As normally used, the projector throws an image directly on to a full-sized cinema screen; by modifying the final projection lens an image some 4 × 3 in. can be produced at an image distance

of about 30 in. This has to be focused on to the mosaic and once again the geometry of the camera makes an otherwise extremely simple problem very much more awkward. Except for a difference of scale, an optical arrangement has been used very similar to that in figure 4. Two large diameter achromatic lenses act as field lenses in or near to the plane of the first image, after which a camera lens of suitable aperture and focal length forms a second image of the correct dimensions on the mosaic. It is important that an image of the mirror disc of the projector should be formed in the plane of the final camera lens, otherwise a vignetting effect is introduced which causes a small periodic variation in illumination on the mosaic. This is in turn manifested in the television picture as a stroboscopic flicker produced between the rotating mirror disc and the scanning beam of the electron camera.

§ 4. RECEIVER SYSTEMS

In all the television receivers on the market at the time of writing (June, 1938), the picture is formed on the fluorescent screen at the end of a cathode ray tube. This can, of course, be viewed directly without the aid of optical devices of any kind and the only purpose for which lenses might be required would be to enlarge the picture to a size suitable for the number of persons in the audience. There are two methods that can be adopted. In the first, a large lens may be placed a short distance in front of the cathode tube to form an enlarged virtual image of the screen on the ordinary "magnifying glass" principle. This method is only of limited application. There is no difficulty in producing a satisfactory image—distortion is the only aberration of any magnitude that needs watching—but as soon as the lens is introduced, the angle within which the audience can still see the complete picture without any vignetting is limited by the diameter of the lens. With the cathode tube alone, the picture can be viewed from a very oblique angle; hence the increased size can only be obtained at the sacrifice of viewing angle. Again, as the picture will normally be viewed from a distance of several feet, the apparent size of the enlarged picture cannot exceed the apparent size of the lens. It then becomes a question to decide whether it is economically better to use a cathode ray tube with, say, a 7 in. end and magnify the picture with a 14 in. diameter lens or to install a cathode ray tube with a 14 in. end and no magnification. Among the factors that have to be considered are the relative costs of the lens and the tube, the cost of replacing the small or the large tube, the space occupied by the two systems, the relative angles of view and so on.

The other method of enlarging the picture is to throw an image of the end of the cathode tube on to a screen with a projection lens of suitable aperture and field. The success of this method depends primarily on the development of a projection cathode ray tube which is capable of producing a very bright, well-defined image on the fluorescent screen. But it is also imperative that the optical system should have the maximum possible efficiency. This calls for a wide-aperture lens and a screen that makes the utmost use of the light falling on it. A lens with an aperture of f 1·5 can be used, but it is doubtful whether any further increase would be justified in view of the delicate focusing and squaring-on that would be called for. The actual focal length and angular diameter of field to be covered by the lens

depends almost entirely on the size of the cathode tube that can be made, the size of the final picture and the distance of throw required, the proportion of the cost of the receiver that can be allocated to the optical system and so forth. Cost is a vital consideration. A market may be found for sets giving definition of not quite first class quality, in which case there may be a demand for a projection lens of slightly smaller aperture or of maximum aperture but somewhat inferior definition. The type of lens required is likely to be different from those employed in cinema projectors, to the extent that the image on the cathode screen will almost certainly be larger than the size of frame of standard cinema film, although a proportionate increase in focal length will not in normal circumstances be desirable. Much will depend, of course, on whether the picture is to be shown in an average sized living room, in a small hall or lounge, or in a cinema or theatre.

The colour correction of the lens may also call for some attention. A white image is to be preferred, in which case normal visual achromatism should be called for. On the other hand, it may be found that the brightest image can be produced in the cathode tube using a fluorescent material which has a fairly strongly marked colour, such as willemite green. In that case it may be possible to dispense altogether with any colour correction or at least aim only at a partial correction, and this added freedom may then be used to provide a cheaper lens or one with a higher aperture.

Given plenty of light, the most satisfactory screen is a white diffusing surface; this gives a uniformly bright picture and a picture that is equally bright from whatever angle the screen is viewed. On the other hand, it is very wasteful of light, only 12 per cent of the total light scattered by the screen being radiated within a cone of 20° on either side of the normal. In a great many cases the television audience would be seated inside that cone, hence a screen which radiated most of its light within that angle would, for television purposes, be noticeably more efficient than the perfectly diffusing screen. Unfortunately a directional screen of this type generally tends to give a picture in which the centre is considerably brighter than the edges. A directional screen with no central glare spot is one of the needs at the present time for projection television. It may prove to be only a transient need; when a sufficiently bright picture can be generated in the cathode tube, the problem will have disappeared.

§ 5. ACKNOWLEDGEMENT

In conclusion the writer's thanks are due to Messrs Electric and Musical Industries, Ltd., for permission to publish details of the optical systems used in the television transmitting equipment supplied by them to the B.B.C.

REFERENCES

(1) Myers, L. M. *Television Optics* (Pitman 1936). (This book includes a list of references to original papers and patents up to 1936.)
(2) *Television*, **10**, 460 (Aug. 1937).
(3) Goddard, M. J. "The Mihaly-Traub television receiver." *J. Television Soc.* ser. 2, **2**, 235 (1937).

SPECTROSCOPY

Edited by W. JEVONS, D.Sc., PH.D.

I. ATOMIC SPECTRA

BY PROFESSOR A. G. SHENSTONE, PH.D.
Palmer Physical Laboratory, Princeton University

§1. STANDARDS AND MEASUREMENTS

THE visible and near ultra-violet regions of the spectrum are now fairly adequately covered by excellent standards of wavelength. They are being added to yearly. For example, Humphreys at the Bureau of Standards has recently made interferometer measures of the ultra-violet lines of Ne, A, Kr[101]. The infra-red is still rather lacking in standards but it presents no essential difficulties.

The more difficult problem of the establishment of standards in the region below λ2250 has been the object of several researches in recent years. The ordinary method of measurement with a Fabry-Perot etalon is scarcely applicable in this region, but it is possible that the reflection echelon may be usable into the upper wavelength range. Attempts to use it have been made and one note has appeared which indicates that some results may be obtained[102]. The second method is the usual one of overlapping orders. This method has established reasonably good standards all the way into the soft x-ray region[103]. It is, however, always subject to errors and cannot be trusted unless measurements from several different orders and gratings are averaged. The third method depends on the use of the combination principle. If the levels of a spectrum can be determined with great accuracy from measurements in the visible and near ultra-violet, there will usually be lines of short wavelength whose values can be computed accurately. The calculable lines of Cu II have been used as ultra-violet standards, but they suffer from the disadvantage that most of them are excited only in the Schüler tube. Recently Dobbie[23] has published a very complete analysis of Fe II and Green[22] has observed a large number of combinations between Dobbie's terms occurring in the Schumann region betweer λ1960 and λ1550. These lines are emitted by an iron spark in nitrogen and are therefore much more convenient than the Cu II lines as standards; they do not extend very far into the short wavelength region, but there is a possibility that they may be used to fix term values of some of the lighter elements which may then yield shorter lines by a second application of the combination principle.

M.I.T. programme. The programme of wavelength measurement being carried out at the Massachusetts Institute of Technology under the direction of G. R. Harrison

has been accelerated by the construction of two comparators of entirely new design. The radical departure from conventional design was necessitated by the unusual demands put upon these instruments. They must have a travel of over 20 in. and be capable of motion at the rate of about 25 cm. a minute. The elimination of periodic errors from screws of such length presents a very difficult problem, and the lapping of the screws was therefore terminated when the periodic error was reduced to $\pm 3\mu$. The remaining periodic and cumulative errors were then eliminated by a very ingenious three dimensional cam which rotates synchronously with the driving screw. New methods were also devised for the calibration of the screws by effectively converting the comparator into a dividing engine, the scale marks being photographic images of a flashing argon capillary tube instead of diamond scratches[104].

The chief results of the M.I.T. wavelength programme should be mentioned.

(1) The Fe arc has been measured from λ10,000 to λ2000 with 10-metre gratings of 30,000 and 15,000 lines per in. giving dispersions from 0·4 to 3·3 A. per mm. depending on the wave-length region. The number of arc lines has been increased to 10,000 and the number of spark lines to 4000. The accuracy is within $\pm$0·003 A. and many lines are correct to $\pm$0·001 A. as shown by checks with the combination principle.

(2) A new complete description of Ce I and Ce II has been made and an analysis including 2300 Ce II lines carried out using the spectral interval sorter (*Reports*, **4**, 319 (1937)) and the spectral interval recorder[105].

(3) A catalogue has been prepared of wavelengths, wave-numbers, and intensities of the 100,000 lines most strongly emitted by the elements under normal excitation between λ10,000 and λ2000, arranged in order of wavelength.*

(4) A complete card catalogue has been made of all atomic lines listed in the literature. It contains 250,000 entries between λ90,000 and λ100. Re-measurements of these lines and many new lines are now being carried out.

Gratings. In the last few years, Prof. R. W. Wood has had great success with new techniques for ruling gratings. Instead of speculum metal, he now uses aluminized pyrex blanks which do not wear or break the diamond and allow the ruling of 30,000 lines per in. over a 6 in. space. Ghosts are practically a thing of the past and the definition is remarkable. He has found it possible to rule 15,000 lines per in. with such control of the groove form that 85 per cent of the energy in the visible is concentrated in the first order, or 60 to 70 per cent in the second with a different diamond setting, and only 1 per cent in the central image for certain angles. Plane gratings have been ruled $6\frac{1}{2} \times 4\frac{1}{2}$ in. and have been used exclusively for the past two years in the spectrograph of the 100-in. telescope at Mt. Wilson in place of prisms. They are far better in the ultra-violet, and with them the interstellar sodium lines at λ3300 have been photographed for the first time. He has also ruled a remarkable parabolic $6\frac{1}{2} \times 4$ in. grating of only 1·5 metres focus for the Harvard observatory. Large transmission replicas of great brilliancy can now be made. They are being used at Harvard in a new programme on the determination of star colours in clusters.

* This catalogue is now being completed and will be published if 500 advance orders can be obtained. The price will be $12.

§ 2. RECENTLY DEVELOPED SOURCES

In recent years, two types of source for spectroscopic observation have been very highly developed. They are the Schüler or Paschen hollow-cathode tube and the atomic-beam tube. In addition, some remarks will be made about arcs run in pure nitrogen.

Schüler tube. The hollow-cathode tube was originally used by Paschen for the excitation of the spectra of gases and of low-melting-point metals, and was later modified by Schüler into the most powerful source of the excessively fine lines required in the examination of hyperfine structure. The essential parts of the tube are a hollow cathode, usually cylindrical and of diameter from 2 to 20 mm., a rare gas at a pressure of a few mm. and an anode which may be placed anywhere in the tube. Since the discharge takes from 200 to 1000 v., the direct-current source must be of sufficient voltage and power to supply this voltage to the tube plus the voltage across the necessary stabilizing resistances. The tube runs with the discharge mainly inside the cathode, the cathode fall being confined to a very thin layer in the interior. The remainder of the inner space is nearly field-free and thus satisfies the condition for the production of fine lines. The excitation of an added element takes place chiefly by collisions of the second kind with the noble gas ions.

For use in the observation of hyperfine structure such tubes are used at low currents and constructed so that the cathode portion can be maintained at liquid-air temperature to decrease Döppler effect. It has proved, in this form, by far the most successful source for hyperfine structure in general although, as will be described later, it is surpassed by the atomic beam for the solution of special problems.

The use of the hollow cathode for the excitation of spectra for atomic analysis is quite simple. The material, usually in metallic form, is placed inside the cathode, which is heated by the discharge. Practically the whole power applied to the tube is dissipated at the cathode. Depending on the relation between temperature and vapour pressure of the element to be observed, the cathode may be cooled or not. For elements such as the alkalis, the cathode cannot be allowed to rise to a high temperature, but for high-melting-point metals it must be arranged to have good heat insulation so that the discharge may heat it effectively. Unless the rare gas is very pure, the discharge is apt to be unstable.

The amount of excitation obtained from such a tube is rather closely restricted by the ionization potential of the rare gas used. It can, therefore, never be much higher than 24·5 v., the I.P. of helium, and correspondingly lower for the other gases. The emitted spectrum is thus confined to the arc and first spark, but in those two spectra the excitation is such as to favour the higher series terms. It is really in this last point that the value of the hollow cathode lies.

The use of this source for high-melting-point elements has been developed recently at Princeton[60]. With a limited potential available the problem is mainly to increase the amount of power applied to the tube without decreasing the ballasting resistance to the point of instability. The solution of the problem is contained in a circuit devised years ago by Steinmetz. This circuit receives power at constant

alternating potential and delivers power at constant current regardless of the load applied. If such a constant current is fed to an ordinary rectifier, the result will be constant direct current. The actual circuit in use will deliver 4 amp. at 1500 v. but can be used perfectly safely for running ordinary arcs which require low voltages. The 1500-v. terminals can, of course, be short-circuited with safety, but the circuit must never be left open. It can easily be seen that such a circuit is ideal for operating any unstable circuit such as a gas discharge. It is effectively like using an infinite applied potential with an infinite resistance in series.

With this circuit applied to a tube in which the cathode was a thin-walled graphite cylinder containing iron suspended by tungsten wires in the centre of a 2-litre pyrex flask, it was found possible to use as high as 3 kw. (2·5 amp. at 1200 v.). With such power, the cathode reaches a temperature at which the vapour pressure of iron is sufficient to partially suppress the helium lines. The spectrum in the visible and near ultra-violet contains a multitude of new lines and in the Schumann region is almost entirely new. With a similar tube platinum has been excited with about 2 kva.

Atomic beam. Although the liquid-air-cooled Schüler tube produces lines which are in most cases fine enough for hyperfine structure analysis, there are problems where still narrower lines are necessary. The atomic beam has supplied the need for such a source. The principle is essentially as follows. By means of restricting slits a narrow beam of atoms is selected from the number evaporating from a liquid surface in a high vacuum. If the beam is narrow and the density so low that collisions between atoms are very improbable, there will be only a very small component of velocity at right angles to the beam. If the depth of the slits in the line of sight be a and the distance between them l, the effective temperature for perpendicular observation is reduced to approximately $(a/l)^2$ times the oven temperature. This factor can easily be made to bring the effective temperature into the region of a few degrees absolute. Minkowski and Bruck[106] have analyzed mathematically the shape of line to be expected, and they find that it is very different from the shape given by an ordinary Doppler formula, being much narrower at the maximum and having wider wings. They have tested their deductions on the cadmium red line using Fabry-Perot etalons up to 11 cm. long. They find evidence of hyperfine structure which is of the order of the natural line-breadth.

Jackson and Kuhn have applied the atomic-beam method to the measurement of hyperfine structure in absorption. They used a beam of potassium about 1·5 cm. thick and succeeded in recording the hyperfine structure of the resonance lines in absorption on a continuum consisting of the narrow lines produced by a discharge tube containing potassium. They used a 10 cm. etalon. The lines were observed as pairs with separations of about 0·015 cm^{-1}, a resolution which is considerably beyond the capability of any other source. Other elements have also yielded to this method.

Meissner has greatly extended the usefulness of the atomic-beam source by adapting it for the observation of emission lines. This is accomplished by bombarding the beam with a strong beam of electrons. He was able to resolve the com-

plete hyperfine structure of the resonance line $^2S_{\frac{1}{2}} - {}^2P_{\frac{1}{2}}$ of sodium by means of a 13-cm. etalon. This corresponds to an actual resolving power of about 3×10^6, the separation being 0·0058 cm.$^{-1}$ The measured separation is equal to the Doppler breadth of the line at a temperature of 5° K., which means that the effective temperature is certainly considerably less than that value.

Meissner has also repeated in emission Jackson's observation of the hyperfine structure of the potassium resonance lines and has found that the magnetic moment is positive, in contradiction to the opposite result indicated by the relative intensities of the absorption lines in Jackson's experiment. The Rabi magnetic method had previously been in disagreement with the spectroscopic result.

Arc or spark in nitrogen. A third type of source which has been valuable in the elucidation of arc and low spark spectra in the Schumann region is the arc or spark in pure nitrogen. The reason for this is two-fold. First, nitrogen is quite transparent to at least 1000 A., while oxygen has an enormous coefficient of absorption in this region. Second, low spark and especially arc lines are absent or almost so from the usual vacuum arc or spark sources. Arc-spectra limits are mainly in the near fluorite region and the arc lines below that point are frequently wide owing to auto-ionization. They, therefore, require high-pressure arcs for their production. It is safe to say that any metallic arc will yield new and valuable data in this region, the recorded lines being confined in most cases to the longer wavelengths. An attachment for the purpose has been used in this laboratory [60] for observations on a large number of spectra including several which are being analysed by the Bureau of Standards. As described, the limit of observation is at λ1240 because of the fluorite lens, but we have reached λ1000 by using a very small slit through which the nitrogen leaks into the spectrograph which is exhausted by two hypervac pumps.

It is perhaps worthy of remark that arcs between clean electrodes in pure nitrogen have many forms, one of which is really a glow discharge in which the cathode absorbs most of the energy. In this form the arc may show a sharp division, the anode half emitting the blue spectrum of nitrogen and the other half the spectrum of the metal. In its ordinary form, the metallic arc in nitrogen emits the lines of the twice-ionized atom with considerable intensity.

§ 3. EXCITATION OF INNER ELECTRONS

Most physicists are accustomed to thinking of optical spectra and X-ray spectra as being phenomena of very different character. That such an attitude is incorrect has been shown in recent years by H. Beutler in a series of important papers.

We can describe the state of an atom or ion by means of a set of symbols which tell us the quantum numbers of the various electrons. For example, the zinc atom has a lowest energy-state described by $1s^2 2s^2 2p^6 3s^2 3p^6 3d^{10} 4s^2$. Ordinary states of the atom are then due to the excitation of one or both of the 4s electrons ($4s\,np$, $4p^2$, etc.). If we pass a suitable continuous spectrum through the vapour we should observe line-absorption corresponding to every allowed transition from the ground state and these absorption lines will form a series ($4s^2 - 4s\,np$) converging to the

point in the spectrum corresponding to the energy necessary for the complete removal of one 4s electron. This in the past has usually been considered as the whole atomic line-absorption spectrum. But if one considers the whole of the electron structure instead of just the outer two electrons, it is obvious that many other absorption lines are possible. For instance, considering the outer 12 electrons, the structure $3d^9 4s^2 np$ has correct parity to combine with $3d^{10} 4s^2$ and there should therefore be several series of absorption lines converging to a point corresponding to the energy of removal of a 3d electron. This is known from the spark spectrum to be at 140,000 cm.$^{-1}$ Such series have not been observed in the past simply because of the lack of a suitable continuum, a lack which no longer exists. Using Hopfield's He_2 continuum, which extends from $\lambda 1000$ to $\lambda 600$, Beutler has observed in Zn the series of lines $3d^{10} 4s^2 {}^1S_0 - 3d^9 4s^2 np\ {}^3P^o_1$, ${}^1P^o_1$ ${}^3D^o_1$ and possibly $3d^{10} 4s^2 {}^1S_0 - 3d^9 4s^2 nf$ ${}^3P^o_1$, ${}^1P^o_1$. But there is no reason why we should confine ourselves to a consideration of the 3d and 4s electrons. The excitation of any other electron will lead to similar results. This means that the x-ray absorption spectrum of an element in the gaseous state should consist of a set of absorption edges, each one of which has on its long wavelength side a series of absorption lines. An indication of such lines has been found by D. Coster in argon. The difficulty of observation arises from the very narrow region in which such a series lies.

It should be remembered that when the zinc atom is raised by absorption of radiation to a discrete state such as $3d^9 4s^2 np$, it has been given far more than enough energy to ionize it. In other words, as far as energy is concerned, the atom may exist either in a discrete state or in the form of an ion ($3d^9 4s^2$) plus an electron with kinetic energy. This possibility of two different states of the atom with identical energy must be considered with some care.

The energy of the atom in its state of ion-plus-electron is not quantized but forms a continuum extending from the point where the discrete energy states cease, i.e., from the ionization point. One can consider the continuum as being an actual indefinite extension of the series of discrete levels. This means that the quantum characteristics of the continuum are identical with those of the series, and that, therefore, there is just one continuum for each series. For example, one can characterize the continua of an alkali as ${}^2S_{\frac{1}{2}}$, ${}^2P^o_{\frac{1}{2}}$, ${}^2P^o_{1\frac{1}{2}}$, ${}^2D_{1\frac{1}{2}}$, etc., continua. Now, if an atom finds itself in a discrete state above the ionization point, it can ionize itself automatically with no change in its angular momentum vectors if there is a continuum with exactly the same characteristics. This can happen in the strict Russell-Saunders case only if the L, S, J and parity are the same. In other cases, the rules will be different, but as far as I know have never been worked out theoretically. The reverse effect is also possible, in which a state of the continuum changes automatically to a discrete state. It is evident in the first case, usually called auto-ionization, that the average life-time of the discrete state will be more or less reduced depending on the probability of the auto-ionization. For instance, in a case of intermediate coupling, a level may have the characteristics of several different types of term to a greater or less degree. In that case its probability of auto-ionization will depend upon the admixture in it of those types which can find a continuum.

This probability will in turn determine the width of any line originating in the discrete state. The effect on the line-width is equivalent to the classical case where line-width is dependent on the finite length of the wave-train. To use Zn again as an example, the continua have the characteristics of ^2S, ^1S, $^3P^o$, $^1P^o$, ^3D, ^1D, $^3F^o$, $^1F^o$ etc. The series of discrete levels found by Beutler are due to the structure d^9s^2p giving $^3P^o\,^3D^o\,^3F^o$, $^1P^o\,^1D^o\,^1F^o$. Of these levels $^1P^o_1$, $^3P^o_1$, $^3D^o_1$ can combine with the ground state to give absorption. The first two find continua but the third does not, with the result that there are two series of diffuse lines and one of sharp. If the coupling were far from Russell-Saunders, the three levels would probably all auto-ionize.

The probability of auto-ionization is also closely connected with the probability that a series will be perturbed. If a series of levels based on a higher limit shows auto-ionization and also includes levels in the discrete part of the level diagram, these latter levels will perturb the ordinary series. For instance, in Hg I the series $d^{10}sp\ ^1P$ is strongly perturbed by the level $d^9s^26p\ ^1P$, and the level $d^9s^27p\ ^1P$ which lies above the series-limit is very diffuse from auto-ionization.

Further complications which occur arise from the fact that the limits involved may be multiple terms or widely separated terms due to a single structure. In either case the coupling may be $L-S$ or $J-j$ or of any intermediate type. In such cases it is very difficult to predict whether or not a given series will show auto-ionization. As was remarked above, this problem has not been sufficiently considered theoretically.

The simplest case of this sort has been observed by Beutler in the spectra of the heavy rare gases. In these arc spectra the limit is $p^5\,^2P^o$, and the $^2P^o$ interval becomes very great in krypton and xenon. Certain of the ordinary series converge to the lower limit $^2P_{1\frac{1}{2}}$, but others must converge to the upper limit $^2P_{\frac{1}{2}}$ and should therefore have many levels above $^2P_{1\frac{1}{2}}$ which have never been detected in emission spectra. The reason is obvious from Beutler's observations. Such series appear with striking intensity in absorption, but the lines are very diffuse and unsymmetrical, owing undoubtedly to auto-ionization. The life-time of the levels is too short for them to produce emission lines in discharges at the low pressures which are normally used.

In Kr I the series $p^5(^2P_{\frac{1}{2}})ns$ and $p^5(^2P_{\frac{1}{2}})\,nd$ start below the limit $p^5\,^2P_{1\frac{1}{2}}$. In that region the series-lines are very weakly absorbed, but after passing that limit the probability of absorption increases by a factor in the neighbourhood of 100. This is a very remarkable fact and is, as yet, inadequately explained. The same effect occurs in other cases, but it is most striking in the rare gases. It would be interesting to know whether a similar increase in absorption probability occurs whenever a limit is passed. This could be tested in the case of Cs except for the fact that there are so many absorption lines that it is impossible to be quite sure of the series relationships. The series, in this case, converge to four widely spaced limits $^3P_2$, $^3P_1$, $^3P_0$, $^1P_1$.

This very great absorption between the two series limits accounts for the fact that the series-limit term in the dispersion formulae of the rare gases increases in importance as we go to the heavier gases. The heavier the gas, the greater the distance between the limits and therefore the more series terms fall between them. The series-limit term in the dispersion formula is much more important than the term due to

the resonance lines, and it now appears to be due to line absorption rather than to absorption in the continuum.

Beutler's observations[107] have included the following atoms. The new levels found are given in each case.

Hg	$5d^9 6s^2 np\ {}^3P^o_1, {}^1P^o_1, {}^3D^o_1$		$n=6-18$
	$nf\ {}^3P^o_1, {}^1P^o_1$		$n=5-10$
Cd	$4d^9 5s^2 np\ {}^3P^o_1, {}^1P^o_1, {}^3D^o_1$		$n=5-14$
	$nf\ {}^3P^o_1, {}^1P^o_1$		$n=4-7$
Zn	$3d^9 4s^2 np\ {}^3P^o_1, {}^1P^o_1, {}^3D^o_1$		$n=4-12$
	$nf\ {}^3P^o_1, {}^1P^o_1$		$n=4-5$
K	$3p^5 4s^2\ {}^2P^o$		
Cs	$5p^5 6s ns$	$J-j$ coupling	$n=6-13$
	nd	,,	$n=5-12$
	(116 lines of which 54 are identified)		
Rb	$4p^5 5s ns$	$J-j$ coupling	$n=5-8$
	nd	,,	$n=4-7$
Tl	$6s6p np$	,,	
	$5d^9 6s^2 6p np$	,,	
A	$3p^5({}^2P)\ ns$	,,	$n=4-8$
	nd	,,	$n=3-16$
Kr	$4p^5({}^2P)\ ns$	,,	$n=5-12$
	nd	,,	$n=4-14$
Xe	$5p^5({}^2P)\ ns$	,,	$n=8-12$
	nd	,,	$n=6-18$

In all cases Beutler finds that the positions of the new levels with respect to their proper limits show a close correlation to the equivalent levels of the following arc spectrum rather than to the levels of the ordinary spectrum of the same element. For example, the levels $3d^9 4s^2 nl$ of Zn I have values much nearer to the levels $3d^{10} 4s^2 nl$ of Ga I than to the levels $3d^{10} 4s nl$ of Zn I. This is to be expected since the outer structure should play a larger role in the determination of the energy differences than does the inner structure. It should be pointed out, however, that Beutler places too much reliance on Hund's predictions regarding the correlation of series to component limits. As it is given in the 1927 edition of Hund's *Linienspektren*, this correlation is definitely not applicable in the case of inverted terms. In fact, it is hardly possible to predict at all to what limits component series will converge. This introduces some uncertainty into the details of Beutler's analysis. The case of Cs is the most complicated one observed. Not only are there four limits ($p^5 s\ {}^3P$, ${}^1P$), but since lines are observed beyond the last of these, there must be absorption series due to double-electron excitation.

It should be noticed that Beutler's method can determine the values of only such states as can combine with the lowest level of the atom. There must be, of course, a multitude of levels which have other characteristics. Such levels cannot be found from absorption observations and may be very difficult to observe in emission because of the effects of auto-ionization. The effects of auto-ionization can, however, be partially eliminated by using sources at higher pressures and with high electron densities. Such conditions produce the reverse effect, i.e. they favour transitions from the continuum into the discrete states. This has been shown experimentally in the case of the transition $d^9ss \rightarrow d^9sp$ in Cu I[108]. We have observed in this laboratory that ordinary arcs, run in nitrogen (for transparency), do emit lines of great width and intensity in the Schumann region beyond the ordinary series limit. We have observed lines whose widths are of the order of 100 to 1000 cm^{-1}. Lines of this type do not appear at all in a low-pressure source.

Since Beutler's work has now proved experimentally the existence of arc levels corresponding to inner-electron excitation, we should expect to find under proper conditions whole new systems of lines in almost any element. For instance, in the alkali metal spectra there should be found the lines $p^5s^2\,^2P - p^5sp\ ^{4*2*2}SPD$, and their position should be near the position of $p^6s^2\,^1S - p^6sp\,^3P^o, ^1P^o$ in the following alkaline earth. The conditions for the observation of such lines in the laboratory are probably high pressures and currents. Helium is a special case. It has no inner shells for excitation, but it should possess a secondary arc spectrum due to two-electron excitation. In fact, Kruger believes that he has observed two lines in the far ultra-violet which involve the levels $2s^2\,^1S$ and $2s\,2p\,^3P$.

A considerable number of spectroscopists believe that the unsolved problem of the coronal lines* may be closely connected with the type of spectra just discussed. The light of the solar corona contains about 60 emission lines of no known origin and no lines at all of any known spectrum. We can hardly doubt that these lines are due to some common known element or elements brought to emission by some very special process which suppresses completely their usual spectra. The fact that the lines are diffuse although the source is of very low density suggests auto-ionization. Because helium is the most likely element to exist in the corona, attempts have been made to observe in the laboratory the double-electron excitation. At low pressures it is possible to produce electron velocities of sufficient magnitude to produce such excitation, but no new lines have been observed[109]. An attempt has also been made using atmospheric pressure and a very high-current arc; this also produced no new lines[110].

§4. ANALYSIS

It is a curious fact that the chief key to the analysis of complex spectra was in the hands of physicists as early as 1901, but no use was made of it until twenty years later. In that year C. P. Snyder published in the Astrophysical Journal a fairly complete numerical analysis of the arc spectrum of rhodium. Nearly 500 lines were fitted into the sort of array of equal frequency differences with which spectroscopists

* See also p. 243.

are now so familiar. He also discovered such arrays in nickel, ruthenium, osmium, iridium, and platinum; they were never published but were found, I believe, amongst Prof. Rowland's papers. It was unfortunate that Snyder chose such very complex spectra, because it was probably the extreme complexity that prevented other physicists from trusting his results to the extent of trying to interpret them. It was not until 1922, when Catalàn discovered multiplets in manganese and Gieseler analysed chromium, that it was realized that the discovery of constant frequency differences, i.e. the Ritz combination principle, was the real key to spectrum analysis. The activity of the succeeding years was so great that the scramble for priority resulted in analyses that were frequently very superficial, if not actually incorrect. Thoroughness was sacrificed to speed. The justification for this superficiality is, of course, that the results were adequate to produce the great monument to spectroscopy—the wave mechanics, which could hardly have been conceived without the great mass of spectroscopic data so rapidly put together.

The line spectroscopists have two fields of research which are more or less separated by matters of technique. The first of these is concerned chiefly with the arc and low spark spectra, and the second with high spark spectra. The two fields will be considered separately.

Arc and low spark spectra. In this field, the problems are the completion of the already partially analysed spectra and the initiation of analysis in the special field of the very complicated spectra, of which the rare earths are a conspicuous example.

To complete an already partially analysed spectrum is much more difficult than to begin a new one, because it is always the easy part that is already done. It is first necessary to make new observations of the spectrum with great thoroughness and accuracy. Accuracy requires large dispersion and resolving power which are normally inconsistent with observation of the faint lines. The latter being necessary, they must be observed on faster instruments with less accuracy. We have, in this laboratory, exposed plates on a Hilger E 1 to the point where the strongest lines are surrounded by more than 100 diffraction fringes, in order to bring up the faint lines which tie together a complete analysis.

For low ionizations the chief sources are the ordinary arc and spark, but they must usually be supplemented by other sources. For example, if one wishes to bring up high series terms in either an arc or first-spark spectrum, the Schüler tube is almost indispensable. Its use with various rare gases always gives useful information, and, especially in spark spectra, may yield the major part of the data. In special cases, other sources such as the low-voltage arc, the exploded wire, the absorption tube, the furnace, the atomic beam may add materially to the required knowledge. If the spectrum is one in which there are levels which auto-ionize, it may be necessary to photograph arcs running at more than atmospheric pressure. Special methods may be required in order to eliminate bands, the bane of the line spectroscopist.

Naturally the spectrum must be observed over as wide a range of wavelengths as possible. With the new infra-red plates which are available, that part of the spectrum is no longer the stumbling-block that it used to be. The Schumann region

is also necessary and can be observed very readily. The observation of arc lines in that region is, however, rather infrequent because most vacuum spectroscopists are apt to be interested in higher spark spectra. We have found that arcs in pure nitrogen are very useful in this region, because the excitation is low and the stages of ionization can be differentiated by the polarity of the lines. Since few laboratories are equipped for observations over the whole wavelength region, it is frequently necessary to cooperate with other laboratories. For example, when Cu II was completely analysed here, the infra-red observations were made at the Bureau of Standards and the vacuum region was photographed at M.I.T. Since then we have acquired a vacuum spectrograph, and we now assist the Bureau by photographing that region for them.

The means of exciting gas spectra are of course rather different. The usual source is the Geissler tube in which the conditions of excitation may be varied by the gas pressure, the admixture of other gases, and by a control of the density and frequency of the current. The electrodeless ring discharge is an excellent source of higher excitations, and the Schüler tube of low.

In all cases, it is a prime necessity that there should be a reasonably consistent intensity estimate of each line and that the line characteristics should be noted.

The primary analysis of a spectrum is a numerical one which is merely the expression of the Ritz combination principle. And, in fact, in the most complex spectra, very little more than that analysis and a determination of J-values is possible. More can be accomplished, however, if the type of coupling is near one of the extreme cases of Russell-Saunders or $j-j$ coupling. The chief criteria for the identification of levels in either case are: series relationships, intensity estimates and line types, the interval rule, and the Zeeman effect, if obtainable. All of these criteria except the first are quite different for the two extreme cases, but blend into each other for intermediate types. The levels of a spectrum which is near the Russell-Saunders end of the scale can usually be named with some ease, and most spectroscopists believe it is worth while to use the terminology of that type even when the deviations from it are considerable. In cases of overlapping configurations, it may or may not be reasonable to use Russell-Saunders notation depending on the degree to which the levels share characteristics.

To go as far with a spectrum as thorough observations permit is a long and tedious job. It requires a wide knowledge of spectral relationships, but even more a willingness to take time to find every possible term. It is a case where the old adage "If a job is worth doing etc." is quite applicable. That rule is not always obeyed even now. Analyses are still being published which are both incomplete and inaccurate.

Much of the above discussion seems to apply only to complex spectra. It is worth while, therefore, to illustrate its applicability to simple spectra by describing results recently obtained in one of the simplest and most frequently observed spectra, that of neutral sodium. Two papers have appeared on the analysis of this spectrum in the last year. The first is by K. W. Meissner[9] on the diffuse series. By means of the atomic-beam method he has succeeded in resolving for the first time

the 3^2D and 4^2D terms. The separations are 0·0494 and 0·0346 $cm.^{-1}$, and they are inverted as are many others of the alkali doublets. H. E. White gave a possible reason for such inversions based on perturbations by inverted terms far above the limit. This explanation has been dealt with theoretically by M. Phillips[111] who was able to show that it is probably adequate. Meissner has also shown that the accepted wavelengths of one member of the sharp series were in error by over 0·25 A. and that a pair of formerly unidentified Fraunhofer lines is actually due to sodium. The second contribution to Na I, by P. Rood and R. A. Sawyer[5], was the observation of eight members of the $^2D-{}^2F$ series of which only two terms were previously known. These new observations also removed a number of Fraunhofer lines from the unidentified list.

The rare earths present a problem which differs rather radically from all others. The complexity of the spectra is of a higher order and the number of lines obtainable seems to be merely a function of the time of exposure. The accuracy of measurement must therefore be much greater than is ordinarily required. Otherwise the probability of finding by accident a large square array of constant differences is very great indeed. W. E. Albertson and G. R. Harrison have illustrated this fact in their paper[54] on Ce II in which they show that the analysis given by K. Haspas[112] is quite incorrect (see *Reports*, **4**, 322 (1937)). In their own analysis of Ce II, Albertson and Harrison find that the average deviation between calculated and observed wave-numbers is lesst han 0·02 $cm.^{-1}$, indicating an accuracy of wavelength measurement of $\pm 0{\cdot}002$ A. The mechanical means devised by Harrison for sorting out equal differences (*Reports*, **4**, 319 (1937)) are practically indispensable in the most complicated spectra.

That the essentials of so many of the rare-earth spectra are now known is due to the skill and persistency of a few men, W. F. Meggers at the Bureau of Standards, H. N. Russell at Princeton, and W. E. Albertson and G. R. Harrison at M.I.T. The experimental contributions of A. S. King of Mt. Wilson have been quite essential to the analysis of these and many other spectra.

Higher spark spectra. The study of higher spark spectra, which was initiated by R. A. Millikan and I. S. Bowen about 1924, has advanced very rapidly since the introduction of the grazing-incidence grating spectrograph. Such instruments have made possible the photography of optical radiation down into the soft X-ray region (*Reports*, **4**, 314 (1937)). The source for spectra of highly ionized atoms is invariably a spark between metal or charged electrodes, a rough control of the degree of ionization being obtained by variations in the size of the external spark gap and in the capacity, inductance and potential of the electric circuit. Condensers of capacity as high as 0·25 μF. charged to a potential of 100,000 V. have been used for extreme excitation. The spectrograms obtained are usually very dense with lines due to several ionizations, and analysis would rarely be possible except for the powerful aid of the X-ray doublet laws. These laws allow the prediction of approximate doublet separations and line positions, and thus provide sufficient information for the proper choice of the strong lines of a sufficiently simple spectrum. They have been empirically extended to higher multiplicities. The spectrograms are usually standardized

		I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII	XIII	XIV	XV	XVI	XVII	XVIII	XIX	XX
1	H	A 4																			
2	He	A 2	A 1																		
3	Li	A 1	B 2	E 1																	
4	Be	A 2	A 1	D 2	E 2																
5	B	B 2	B 2	A 2	D 2	E 3															
6	C	B 2	A 2	A 2	A 2	D 2	E 3														
7	N	B 2	A 2	A 2	B 2	B 2	E 2														
8	O	B 2	A 2	A 2	A 2	B 2	B 2	E 2													
9	F	C 2	B 2	B 2	B 2	B 2	C 2	C 2													
10	Ne	A 1	A 2	C 2	D 2																
11	Na	A 3	C 2	C 2	C 2	C 2	C 2	D 2	D 2	E 2											
12	Mg	A 6	A 1	C 2	C 2	C 2	C 2	D 2	E 2	E 2											
13	Al	A 2	A 2	A 1	C 2	C 2	C 2	D 2	D 2	E 2	E 2										
14	Si	A 7	B 2	B 2	A 2	E 2	E 2	D 2	E 2												
15	P	C 2	C 2	B 2	B 2	A 2	U 2	U 2	U 2	U 2	U 2	U 2	U 2	E 2							
16	S	A 2	B 2	C 2	C 2	D 2	A 2		E 2												
17	Cl	A 7a	A 7a	C 2	D 2	D 2	E 1	C 65	E 2	E 2	E 8	E 8									
18	A	A 1	B 2	C 2	D 3	E 2															
19	K	A 2	A 1	D 2	D 2	E 2	E 2	E 2	E 2	D 2	E 2	U 2									
20	Ca	A 1	A 1	D 1	D 2	D 2	E 2	E 2	E 2	E 2	D 2	E 2	U 2								
21	Sc	B 1	C 1	C 1	E 2	E 2	E 2	E 2	E 2	E 2	C 2	A 2	E 2								
22	Ti	A 10	A 1	C 1	B 1	E 2	E 2	E 2	E 14		E 2	D 2	B 2	E 2							
23	V	A 11	A 12	D 1	D 1	E 2	E 13	E 2	E 2	E 14		E 2	D 2	C 2	E 2						
24	Cr	B 15	B 7a	D 16	D 16	E 1	E 2	E 13	E 2	E 2		E 2	D 2	B 2	E 2						
25	Mn	B 17	B 18	E 19	D 16	E 20	E 21	D 2	E 2	E 2	E 2			E 2	D 2	C 2					
26	Fe	B 22	B 23	E 22	E 19	D 16	E 20	E 21	D 2	E 2	E 2	E 2			E 2	D 2	C 2				
27	Co	B 24	C 1			E 19	E 24		E 21			E 2				E 2	E 2	D 2			
28	Ni	A 1	C 1				E 19			E 21									E 2		
29	Cu	B 1	A 25	E 26				E 19												E 2	
30	Zn	A 1	B 27	D 28	E 29																
31	Ga	C 1	C 1	C 1	D 1																
32	Ge	B 66	C 1	C 1	C 1	E 30															
33	As	C 2	E 2	C 2	E 2	E 1	E 30														
34	Se	B 2	C 2	C 2	C 2	D 2	E 1	E 30													
35	Br	C 1	E 2	D 31	E 2	E 2	E 32	E 32	E 30												
36	Kr	B 1	B 2	C 2	E 2																

Added in Proof: Rb IX, $E^{(36)}$ and Cr XV, $E^{(113)}$ should be added to the diagram; they have been included in the synopsis.

		I	II	III	IV	V	VI	VII	VIII
37	Rb	•A 33	•D 34						
38	Sr	•A 1	•A 35						
39	Y	•A 37	•B 37	•C 1					
40	Zr	•B 1	•B 1	•C 1	•C 1				
41	Cb	•D 38	•D 38	•E 39	•D 40	C 2			
42	Mo	•E 7a	•E 7a		E 39	E 41	C 2		
43	Ma								
44	Ru	•D 1	•D 1						
45	Rh	•C 1	•D 1						
46	Pd	•B 1	•B 1						
47	Ag	•C 17	C 42	D 43					
48	Cd	•A 1	•C 1	•E 1					
49	In	•B 1	•A 44	•C 1	•E 1				
50	Sn	•B 82	•C 45	•D 1	•D 1	•E 1			
51	Sb	D 82	E 2	•E 1	•D 46	D 46	E 30		
52	Te	•D 2		C 2	D 47	•D 1	•D 47	E 30	
53	I	E 48	E 2	E 49	E 2				E 30
54	Xe	B 50	B 50a	B 50					
55	Cs	•A 1	•C 51	E 2					
56	Ba	A 52	A 35		E 2				
57	La	•B 53	•B 53	•D 2					
58	Ce	E 88	C 54	C 55	•C 56				
59	Pr								
60	Nd								
61	Il								
62	Sm	D 57	C 58						
63	Eu	B 59	D 60						
64	Gd	E 58	E 88						
65	Tb								
66	Dy								
67	Ho								
68	Er								
69	Tm								
70	Yb	D 61	D 61	D 61					
71	Lu	E 61	D 81	E 81					
72	Hf	•C 1	•B 62						

		I	II	III	IV	V	VI	VII	VIII
73	Ta	E 63	E 7a						
74	W	•C 64	D 83						
75	Re	•D 67							
76	Os	•C 68							
77	Ir	•C 68	E 85						
78	Pt	•B 1	B 69						
79	Au	•B 1	C 70						
80	Hg	•A 71	•C 1	D 70					
81	Tl	•A 1	•A 72	•B 1	•E 70				
82	Pb	•B 1	•A 2	•B 1	B 73	•D 74			
83	Bi	•B 75	•C 2	•C 2	D 76	E 77	D 74		
84	Po								
85									
86	Rn	•C 78							
87									
88	Ra	C 2	D 2						
89	Ac								
90	Th			D 79	E 80				
91	Pa								
92	U								

Synopsis

Stage of spectrum	Number analysed		Increase
	Atomic Energy States	This diagram	
I	67	76	9
II	56	71	15
III	43	53	10
IV	32	44	12
V	24	33	9
VI	7	28	21
VII	2	20	18
VIII	—	18	18
IX	—	14	14
X	—	9	9
XI	—	10	10
XII	—	6	6
XIII	—	5	5
XIV	—	4	4
XV	—	4	4
XVI	—	2	2
XVII	—	1	1
XVIII	—	1	1
XIX	—	1	1
Total	231	400	169

by means of the lines of common elements which always appear and which have been measured with accuracy by the method of overlapping orders.

Diagram of analysed spectra. The accompanying diagram has been prepared to give a view of the present state of spectrum analysis. For each spectrum a reference number is given through which it should be possible to trace the previous papers on that spectrum. It is probable that, in the mass of literature, some analyses have been missed, and I wish to apologize here to the authors of such papers.

Each spectrum has been given a *very rough* rating A, B, C, D, E. An A means an almost completely analysed spectrum and E one in which only a little is known. B, C, D are intermediate stages and U indicates an analysis unpublished and not seen by me. A dot indicates that an analysis of the spectrum appeared in Bacher and Goudsmit's *Atomic Energy States*.[1] It is presumptuous on my part to give these ratings without a very detailed examination of all the data, especially when it is remembered that I am including the most simple and the most complex in the same scale. My excuse is that it seems the only way to give a picture of the present state of this branch of physics. I would like to remind the reader that the ratings are *very rough* estimates.

The most striking feature of the diagram is the very large number of analysed spectra of high ionization in the range of atomic number from 1 to 29. *Atomic Energy States*[1] contained no analysis of a spectrum above the seventh, and there are now 75. Almost all of these are due to Prof. Siegbahn's Upsala school, B. Edlén being responsible for the majority. Particularly important are the long sequences iso-electronic with Ne I, Na I, Mg I and Al I. The latter three are particularly simple spectra, but far more than the barest analyses have been made. The Ne I sequence is traced only by the combinations of the ground state with the five levels $2p^5 3s\ {}^3P_1\ {}^1P_1$ and $2p^5 3d\ {}^3P_1\ {}^1P_1\ {}^3D_1$.

The ionization potential of Cr XV, the last analysed spectrum in the Ne I sequence, is over 1000 v. and the observed lines are as short as 18·5 A. It can easily be realized from this that the potency of the spark source must be enormous. H. A. Robinson[113] has shown from the observed excitations that the spark must develop effective temperatures in the neighbourhood of 500,000° K. and electron pressures of the order of 10^7 atmospheres. In the same paper he obtains an explanation of the fact that in any particular vacuum spark only two degrees of ionization appear with great intensity. It is this experimental fact that makes it possible at all to sort out the lines of a particular ionization.

The spectra of the elements up to fluorine are the most completely analysed of any group. The concentration of effort on this part of the table and the Russell-Saunders type of coupling have resulted in analyses which approach perfection. In other parts of the table where p-electron spectra prevail, excellent progress has been made.

The spectra iso-electronic with the more complicated arc spectra are of course of much greater difficulty, but a very large number of them have been partially analysed. One can hardly expect to get much more than the fundamental multiplets, so that these spectra are for the most part rated E.

The individuals who have contributed most to the analysis of the high spark spectra are B. Edlén, J. Soderquist, I. S. Bowen, P. G. Kruger, H. A. Robinson. R. J. Lang of Alberta has been bold enough to tackle some of the very heavy elements.

In the field of lower ionizations, where the data are much more easily obtained, there are still many spectra poorly analysed. That even those marked A can be improved is shown by the recent work on Na I, which would certainly have been an A rating without those additions. The arc spectra of some common elements which emit relatively simple spectra are very badly analysed. For example, the silver arc emits about one hundred unidentified lines ranging from very sharp lines to lines over 100 wave-numbers wide. Cu I and Au I are incomplete, and we must still discover new details of Hg I in spite of the great mass of papers which have been published about it.

There are now very few laboratories which are concerned with complex spectra. W. F. Meggers and C. C. Kiess are undoubtedly the most important contributors in this field, but a considerable amount of work is being done at M.I.T. by W. E. Albertson and G. H. Harrison, and at Princeton, especially by Miss C. Moore and H. N. Russell. There is active cooperation between these three laboratories.

In *Atomic Energy States*[1] the number of analysed spectra is 231. The diagram in this report gives 400, an increase of 169 spectra; and the synopsis shows that the majority of these newly analysed spectra are in the range of higher ionizations. The work done on the lower ionizations has resulted mainly in a very great improvement in the completeness of those analyses.

REFERENCES

(1). Bacher, R. F. and Goudsmit, S. *Atomic Energy States* (1932). [Including supplementary list at end of book.]
(2) Robinson, H. A. and Shortley, G. H. *Phys. Rev.* **52**, 723 (1937). [Includes a collection of references to p-electron spectra.]
(3) Edlén, B. *Vetens. soc., Upsala. Nova acta*, **9**, No. 6 (1934).
(4) Heyden, Maria. *Z. Phys.* **106**, 499 (1937).
(5) Meissner, K. W. and Luft, K. F. *Ann. Phys., Lpz.* **29**, 698 (1937); Rood, P. and Sawyer, R. A. *Astrophys. J.* **87**, 68 (1938).
(6) Meissner, K. W. *Ann. Phys., Lpz.* **31**, 505 (1938).
(7) Kiess, C. C. *Bur. Stand. J. Res., Wash.*, **21**, 185 (1938).
(7*a*) Kiess, C. C. (in progress).
(8) Edlén, B. *Z. Phys.* **100**, 726 (1936).
(9) De Bruin, T. L. *Physica*, **8**, 809 (1936).
(10) Meggers, W. F. and Kiess, C. C. *Bur. Stand. J. Res., Wash.*, **19**, 309 (1937).
(11) Meggers, W. F. and Russell, H. N. *Bur. Stand. J. Res., Wash.*, **17**, 126 (1936).
(12) Meggers, W. F. *J. Opt. Soc. Amer.* **28**, 176 (A) (1938).
(13) Kruger, P. G. and Weissberg, S. G. *Phys. Rev.* **48**, 659 (1935).
(14) Kruger, P. G. and Patten, H. S. *Phys. Rev.* **52**, 621 (1937).
(15) Kiess, C. C. *Bur. Stand. J. Res., Wash.*, **15**, 79 (1935).
(16) Bowen, I. S. *Phys. Rev.* **52**, 1153 (1937).
(17) Paul, F. W. *Phys. Rev.* **52**, 923 (1937); Moore, Miss C. (Princeton, in progress).
(18) Curtis, C. W. *Phys. Rev.* **53**, 474 (1938).
(19) Kruger, P. C. and Gilroy, H. T. *Phys. Rev.* **48**, 720 (1935).
(20) Bowen, I. S. *Phys. Rev.* **47**, 924 (1935).

(21) CADY, W. M. *Phys. Rev.* **43**, 322 (1933).
(22) GREEN, L. C. (To appear in *Phys. Rev.* (1939).)
(23) DOBBIE, J. C. *Proc. Roy. Soc.* A, **151**, 703 (1935).
(24) BOWEN, I. S. *Phys. Rev.* **53**, 889 (1938).
(25) SHENSTONE, A. G. *Philos. Trans.* **235**, 195 (1936).
(26) GIBBS, R. C. and VIEWEG, A. M. *Phys. Rev.* **43**, 1092 (A) (1933) and unpublished material.
(27) KALIA, P. N. *Ind. J. Phys.* **10**, 463 (1936).
(28) MAJUMDAR, K. C. *Ind. J. Phys.* **10**, 178 (1936).
(29) BLOCH, L. and E. *Ann. Phys., Paris*, **5**, 325 (1936).
(30) KRUGER, P. G. and SHOUPP, W. E. *Phys. Rev.* **46**, 124 (1934).
(31) RAO, K. R. and KRISHNAMURTI, S. G. *Proc. Roy. Soc.* A, **161**, 38 (1937).
(32) RAO, A. S. and K. R. *Proc. Phys. Soc.* **46**, 163 (1934).
(33) MEISSNER, K. W. *Z. Phys.* **98**, 353 (1935–6).
(34) LAPORTE, O., MILLER, G. R. and SAWYER, R. A. *Phys. Rev.* **38**, 843 (1931).
(35) SAUNDERS, F. A., SCHNEIDER, E. G. and BUCKINGHAM, E. *Proc. Nat. Acad. Sci., Wash.*, **20**, 291 (1934).
(36) MACK, J. E. *Phys. Rev.* **38**, 193 (1931).
(37) HO, I-DJEN and SAWYER, R. A. *Phys. Rev.* **51**, 1020 (A) (1937).
(38) MEGGERS, W. F. and SCRIBNER, B. F. *Bur. Stand. J. Res., Wash.*, **14**, 628 (1935).
(39) ELIASON, A. Y. *Phys. Rev.* **43**, 745 (1933).
(40) LANG, R. J. *Zeeman Verhandelingen*, **44** (1935).
(41) TRAWICK, M. W. *Phys. Rev.* **48**, 223 (1935).
(42) GILBERT, W. P. *Phys. Rev.* **47**, 847 (1935).
(43) GILBERT, W. P. *Phys. Rev.* **48**, 338 (1935).
(44) PASCHEN, F. and CAMPBELL, J. S. *Ann. Phys., Lpz.*, **31**, 29 (1938).
(45) MCCORMICK, W. W. and SAWYER, R. A. *Phys. Rev.* **52**, 249 (1937).
(46) BADAMI, J. S. *Proc. Phys. Soc.* **43**, 538 (1931).
(47) RAO, K. R. *Proc. Roy. Soc.* A, **133**, 220 (1931).
(48) PRICE, W. C. *Phys. Rev.* **48**, 477 (A) (1935).
(49) SETH, J. B. *Nature, Lond.*, **127**, 165 (1931).
(50) BOYCE, J. C. *Phys. Rev.* **49**, 731 (1936);
HUMPHREYS, C. J. *Bur. Stand. J. Res., Wash.*, **16**, 639 (1936).
(50*a*) HUMPHREYS, C. J. *Phys. Rev.* **53**, 490 (A) (1938).
(51) OLTHOFF, J. and SAWYER, R. A. *Phys. Rev.* **42**, 766 (1932).
(52) MEGGERS, W. F. *Bur. Stand. J. Res., Wash.*, **10**, 669 (1933).
(53) RUSSELL, H. N. and MEGGERS, W. F. *Bur. Stand. J. Res., Wash.*, **9**, 625 (1932).
(54) HARRISON, G. R. and ALBERTSON, W. E. *Phys. Rev.* **53**, 940 (A) (1938).
(55) RUSSELL, H. N., KING, R. B. and LANG, R. J. *Phys. Rev.* **52**, 456 (1937).
(56) LANG, R. J. *Phys. Rev.* **49**, 552 (1936).
(57) ALBERTSON, W. E. *Phys. Rev.* **52**, 644 (1937).
(58) ALBERTSON, W. E. *Phys. Rev.* **49**, 208 (1936).
(59) RUSSELL, H. N. and KING, R. B. *Phys. Rev.* **46**, 1023 (1934).
(60) ALBERTSON, W. E. *Phys. Rev.* **45**, 499 (L) (1934).
(61) MEGGERS, W. F. and SCRIBNER, B. F. *Bur. Stand. J. Res., Wash.*, **19**, 651 (1937).
(62) MEGGERS, W. F. and SCRIBNER, B. F. *Bur. Stand. J. Res., Wash.*, **13**, 625 (1934).
(63) KIESS, C. C. and H. K. *Bur. Stand. J. Res., Wash.*, **11**, 277 (1933).
(64) LAUN, D. D. *Bur. Stand. J. Res., Wash.*, **21**, 207 (1938).
(65) PHILLIPS, L. W. *Phys. Rev.* **53**, 248 (1938).
(66) KIESS, C. C. *Phys. Rev.* **53**, 931 (1938).
(67) MEGGERS, W. F. *Bur. Stand. J. Res., Wash.*, **10**, 757 (1933).
(68) ALBERTSON, W. E. *Phys. Rev.* **53**, 940 (A) (1938).
(69) SHENSTONE, A. G. *Philos. Trans.* **237**, 453 (1938).
(70) MACK, J. E. and FROMER, M. *Phys. Rev.* **48**, 357 (1935).
(71) MURAKAWA, K. *Z. Phys.* **108**, 168 (1938).
(72) ELLIS, C. B. and SAWYER, R. A. *Phys. Rev.* **49**, 145 (1936).

(73) CRAWFORD, M. F., MCLAY, A. B. and CROOKER, A. M. *Proc. Roy. Soc.* A, **158**, 455 (1937).
(74) SCHOEPFLE, G. K. *Phys. Rev.* **50**, 538 (1936).
(75) CLEARMAN, H. E. *Phys. Rev.* **53**, 941 (A) (1938).
(76) MCLAY, A. B. and CRAWFORD, M. F. *Phys. Rev.* **44**, 986 (1933).
(77) SCHOEPFLE, G. K. *Phys. Rev.* **47**, 233 (1935).
(78) RASMUSSEN, E. *Z. Phys.* **80**, 726 (1933).
(79) LANG, R. J. (In progress.)
(80) LANG, R. J. *Phys. Rev.* **49**, 416 (A) (1936).
(81) MEGGERS, W. F. and SCRIBNER, B. F. *Bur. Stand. J. Res., Wash.*, **19**, 31 (1937).
(82) MEGGERS, W. F. (In progress.)
(83) LAUN, D. D. *Bur. Stand. J. Res., Wash.*, **21**, 207 (1938).
(84) MOORE, Miss C. (Princeton, in progress).
(85) SHENSTONE, A. G. (Princeton, in progress).
(86) ALBERTSON, W. E. (*M.I.T.*, in progress).
(101) HUMPHREYS, C. J. *Bur. Stand. J. Res., Wash.*, **20**, 17 (1938).
(102) MACADAM, D. L. *Phys. Rev.* **50**, 185 L (1936).
(103) EDLÉN, B. *Z. Phys.* **85**, 85 (1933).
(104) HARRISON, G. R. *Rev. Sci. Instrum.* **9**, 15 (1938).
(105) HARRISON, G. R. *J. Opt. Soc. Amer.* **28**, 290 (1938).
(106) MINKOWSKI, R. and BRUCK, H. *Z. Phys.* **95**, 274 (1935).
(107) BEUTLER, H. Papers starting with *Z. Phys.* **86**, 495 (1933) and extending to *Z. Phys.* **93**, 177 (1935).
(108) ALLEN, C. W. *Phys. Rev.* **42**, 39 (1932).
(109) BUNDY, F. P. *Phys. Rev.* **52**, 452 (1937).
(110) SHENSTONE, A. G. *Phys. Rev.* **53**, 941 (A) (1938).
(111) PHILLIPS, M. *Phys. Rev.* **44**, 644 (1933).
(112) HASPAS, K. *Z. Phys.* **96**, 410 (1935).
(113) ROBINSON, H. A. *Z. Phys.* **100**, 636 (1936).

II. HYPERFINE STRUCTURE

BY M. F. CRAWFORD, PH.D.
McLennan Laboratory, University of Toronto

§1. INTRODUCTION

The progress during the past two years has continued along the general lines described in previous *Reports*[1], namely, the determination of nuclear mechanical (I) and magnetic (μ) moments, quadrupole moments, and isotope shifts. The experimental methods that have yielded practically all the results in this period are the spectroscopic analysis of hyperfine structure and the magnetic deflection of atomic beams. The principles of these methods have been described[1], so only the improvements in the experimental technique, the new data, and the interpretation of the data will be reviewed here.

§2. SPECTROSCOPIC METHODS

The spectroscopic methods, through the development of atomic-beam sources, can now be used to measure directly very small hyperfine separations of the order of 0·003 cm^{-1}, which previously could be determined only by magnetic deflection

methods and polarization of resonance radiation. In Jackson and Kuhn's[2] source the radiation emitted by the atoms of an element is passed through a collimated atomic beam of the same element, in a direction perpendicular to the beam, before falling on the slit of the spectrograph. As the atoms in the collimated beam have only very small components of velocity in the direction of observation, the structure of a resonance line appears as very narrow absorptions on the background of the emission line. An atomic beam is also utilized by Meissner and Luft[3], but in their source the beam is excited by collisions with an electron beam. The emission is observed in a direction perpendicular to the two beams.

The widths of the hyperfine components can be reduced in both sources by collimation to correspond to the Döppler widths characteristic of a source at a temperature of a few degrees Kelvin. Jackson and Kuhn[2] for K and Meissner and Luft[3] for Na have measured directly with a Fabry-Perot interferometer separations which correspond to practical resolving powers of $3{\cdot}3$ and $2{\cdot}9 \times 10^6$, respectively. In both cases the components were separated, so that the limit of resolution probably exceeded 5×10^6. This was attained by using plate separations up to 13 cm. and by enclosing the interferometer in a constant-temperature air-tight box.

The use of the atomic beam in absorption is confined to those lines involving the normal levels of the atom that show a measurable absorption, in general, the resonance lines. Thus the hyperfine separations of just a few levels of the neutral atom can be determined. Although these levels are usually suitable for calculating μ, they frequently are not the best choice for determining I. However, I can be determined from either the relative intensities or the Zeeman effect of the hyperfine components. The Zeeman effect, when sufficiently resolved, gives I uniquely; whereas relative intensities which are difficult to measure accurately are rather ambiguous for large values of I. Both procedures have been used by Jackson and Kuhn. In contrast, more lines can be studied with the atomic beam in emission, and those most suitable for determining I can be selected. Only lines of the arc spectra have been investigated as yet, but Meissner mentioned that the strong Mg II lines were visible under proper excitation conditions. The emission, however, is relatively weak and exposures of several hours are necessary, whereas in absorption exposures of only a few minutes are required.

To date, hyperfine structures have been measured, using the atomic-beam source, in lines of Na, Mg, Al, K, Rb, Ag, in absorption; and Na, Mg, K, in emission. The separations of the ^2S levels of the alkalis are in good agreement with those obtained by the magnetic deflection method; and the accurate separations of the ^2P levels provide additional data for the calculation of the magnetic moments. The hyperfine structures of both isotopes of Ag have been resolved and their nuclear moments determined. The structures observed in Mg lines are of isotopic origin, and the complete resolution of the components in emission has removed an apparent anomaly in earlier observations. In the case of Al, however, there is a discrepancy between the value of I determined from Al I lines by intensity measurements[4] and the value obtained by the interval rule from suitable levels of Al II[5]; the discordance is probably due mainly to the insensitivity of the intensity method for large I values.

The results are quite promising and indicate that it will be possible to determine the moments of a number of nuclei for which, as yet, no hyperfine structure has been observed.

New data have also been obtained with the standard types of sources. Space does not permit a discussion of the individual papers, but the scope of the results will be indicated by enumerating the investigated elements in groups that correspond roughly to the type of information obtained:

(1) The spin and, in general, the magnetic moment of at least one isotope of each of the elements Al, Zn, As, Kr, Sr, Rh, Ba, Lu have been determined, in some cases confirming, in others superseding the previous values.

(2) The elements K, Sc, Ga, Rb, In, Re, Pt, Tl have been investigated mainly to obtain new data for the calculation of magnetic moments.

(3) Quadrupole moments have been determined for the elements Ga, As, Kr, In, Xe, Lu, Re.

(4) Isotope shifts have been measured for Li, Ne, A, W, Pt.

In the investigations of Ne[6], A[7], Sr[8] the relative abundance of the rare isotope was increased in the sample used. This made it possible to measure isotope shifts in the weaker lines of Ne and A; with the more complete data relative shifts can be assigned to all the levels of several configurations. In the case of Sr practically a pure sample of ^{87}Sr was used, and its I and μ determined with greater certainty. The new value of I is appreciably larger than the former value, which was uncertain owing to the masking of the ^{87}Sr components by the intense components of the even isotopes. This is an important result since it removes what appeared to be a contradiction of the theory of β-disintegration. As ^{87}Rb has a long half-life, the theory predicts that a large change in I should be associated with the β-disintegration of ^{87}Rb into ^{87}Sr. According to the new value $\Delta I = 3$; the previous value indicated $\Delta I = 0$ or 1.

The Zeeman effects of the hyperfine structures of certain lines of K, Cd, Cs have been investigated with high resolving power. The positions and intensities of the components in these three cases agree with the theoretical predictions within the experimental error.

§3. MAGNETIC DEFLECTION METHODS

Experiments on the magnetic deflection of atomic beams, particularly the method of zero moments for determining I and μ, and the method of non-adiabatic transitions for determining the sign of μ, have already been described[1]. These methods have been used in the past two years to determine the nuclear moments of ^{6}Li, ^{41}K, and the signs of the magnetic moments of ^{6,7}Li, Na, ^{39,41}K, ^{85,87}Rb, Cs. The results are in good agreement with the spectroscopic data, now that the discrepancy for ^{39}K has been eliminated.

There also have been two important developments in this field[9, 10], one an extension of the zero-moment method, the other a new method, which afford a very direct and accurate determination of μ. Previously the magnetic moments of all

nuclei, except ^{1,2}H, had been derived from the hyperfine separations of atomic levels. This was true of the magnetic as well as the spectroscopic methods. The derivation is based on the assumption that the interaction between the nuclear and the electronic angular momentum vectors results from an intrinsic nuclear magnetic moment and in addition a small quadrupole moment. The evaluation of μ involves a calculation of the magnetic field at the nucleus due to the electronic configuration; and as yet, except in a few cases, only semi-empirical formulae are available for the calculation[11]. Although the results certainly confirm the existence of nuclear magnetic moments, the values of μ obtained from the splittings of different levels of an atom usually show a spread of 20 per cent. This undoubtedly is due partly to the approximate formulae and partly to perturbations which, in general, cannot be adequately taken into consideration. There remains, however, the possibility that other interactions not contemplated in the theory may contribute to the hyperfine separations[12]. A direct determination of μ affords a decision on this possibility, and accurate values will critically test the nuclear models proposed to account for the observed moments.

Millman, Rabi and Zacharias[9] have applied the method of zero moments to atomic beams of In under conditions of sufficiently high resolution to allow a measurement of the fine structure of the zero-moment peaks. From the fine structure it is possible to determine μ directly, i.e. independently of the assumption as to the nature of the interaction between the nucleus and the electronic configuration. To elucidate this, one recalls that according to the formula of Breit and Rabi[13] the component of magnetic moment in the field direction (μ_m) for a state with quantum number m, in general, changes with field strength; and for negative values of m there are certain values of the field for which μ_m is zero in two of the $2(2I+1)$ states. Thus for each of these characteristic values of the field the atoms in the two states of zero moment will be undeflected in the inhomogeneous field, and a maximum, or zero-moment peak, will be recorded by a detector at the centre of the beam. The formula of Breit and Rabi, however, is approximate in that small terms of the order of μ have been neglected. The more exact formula[9] shows that the atoms in each of the two states with the same negative m have their zero μ_m at two slightly different values of the field. The difference is such that to a high degree of accuracy the ratio $\Delta H/H$, where ΔH is the interval between the two zero-moment values of the field and H is the average of the two values, is equal to the ratio of the nuclear to the extranuclear magnetic moment, multiplied by a factor involving the quantum numbers. Thus by measuring the fine structure of the zero-moment peaks and knowing the g-factor of the normal level of the atom we can evaluate μ directly.

The results obtained for ^{115}In are: $\mu = 6{\cdot}4 \pm 0{\cdot}2$ nm, $I = 9/2$, confirming the spectroscopic value; and 0·381 cm^{-1} for the splitting of $^2P_{\frac{1}{2}}$ compared to the spectroscopic value 0·380. The value of μ derived from this separation is 6·1, in fair agreement with the above value. The values 5·0 and 5·5 derived from the 6 ^{2}S level of In I and the 5snl levels of In II, respectively, are appreciably lower.

Rabi and his co-workers[10] have also devised a new method of directly measuring μ with unprecedented accuracy. The method depends on non-adiabatic transitions

in a magnetic field. In a theoretical paper[14] Rabi considers the problem of determining the sign and magnitude of μ in cases where the atom in its normal level has no electronic angular momentum. He shows that, for such an atom in a magnetic field rotating about an axis inclined to the field, the probability of a non-adiabatic transition depends on the frequency of rotation, the angle of inclination, and the Larmor frequency, which involves the sign and magnitude of μ. He points out how this dependence can be used to determine μ completely.

In the actual experiments[10] an oscillating field, to which the theory can be adapted, is used. The procedure is to pass a beam of molecules in a $^1\Sigma$ state, first through a weak inhomogeneous field where they are spread out, then through a strong inhomogeneous field, deflecting in the opposite direction, which refocuses certain components of the beam on to a detector. A re-orientating field is placed between the two inhomogeneous fields. It consists of a strong homogeneous field perpendicular to the beam, and a weak oscillating field at right angles to the larger homogeneous field. The homogeneous field is produced by an electromagnet with gap 6 mm. wide and 5 cm. long giving fields up to 6000 gauss. The oscillating field is obtained by passing a current of radio frequency through a loop of wire, in the form of a hair-pin, placed in the gap with its axis parallel to the beam.

If the homogeneous field is sufficiently strong to de-couple completely the nuclear spins from each other and from the molecular rotation, the weak oscillating field will induce non-adiabatic transitions relative to the strong field when its frequency approaches the Larmor frequency of the nuclear angular momentum vector in question. For the molecules that undergo transitions the conditions in the second inhomogeneous field are no longer correct for refocusing and the intensity at the detector drops. The actual procedure is to vary the homogeneous field (H) for some chosen frequency ($\nu = 3{\cdot}5 \times 10^6$ sec.$^{-1}$) of the oscillating field until resonance is observed, by a drop in intensity at the detector and a subsequent recovery when resonance is passed. μ is obtained directly from the value of H for resonance and the known value of ν by the equation $\nu = \mu H/Ih$, where the term on the right is the Larmor frequency with I, the spin, in units $h/2\pi$.

LiCl, LiF and NaF have been investigated and very sharp resonances observed. The common values of ν/H for LiCl and LiF are those of the Li isotopes, and the ν/H common to LiF and NaF is that of F. The values of ν/H obtained from different molecules are concordant within a few parts in a thousand. As the spins of ^{6,7}Li and F are known, the measurements give their magnetic moments directly. They are: 2·79 for ^{1}H[15], 0·823 ± 0·005 for ^{6}Li, 3·265 ± 0·016 for ^{7}Li, 2·635 ± 0·014 nm for ^{19}F. The values derived from hyperfine separations are 2·85 ± 0·15 for ^{1}H; 0·85 for ^{6}Li; 3·20, 3·28, 3·30, depending on the level and the calculation, for ^{7}Li; and about 3, with a range from 1·9 to 3·8, for ^{19}F.

It is noteworthy that the directly determined values, including that of In, agree with those derived from hyperfine separations to within roughly ± 10 per cent. This is the order of the spread of the values obtained from the different levels of an atom, which, for the want of a better criterion, has been used in the past as an estimate of the reliability of the values. It appears, then, that the μ values obtained

from hyperfine separations are certainly of the correct order of magnitude and, in general, are probably reliable within ± 10 per cent. The good agreement between the directly determined moments of ^{1}H and ^{7}Li and the values 2·85 and 3·28, which are the most trustworthy of the moments derived from hyperfine separations, shows that for these two nuclei interactions other than those assumed in the theory certainly are small. This conclusion, however, cannot be extended to other nuclei until more accurate formulae are available for calculating μ from hyperfine separations.

Space does not permit the inclusion of a summary of the data on nuclear moments. Tables of I and μ values, with references, are given in a recent publication by Schmidt[16], and the known quadrupole moments are summarized in a paper by Korsching[17].

§4. NUCLEAR THEORY

It has been pointed out in a previous *Report* (**3**, 373 (1936)) that the proposed vector models of the nucleus can only be considered as preliminary attempts to order nuclear moments. Recently a rational theory of nuclear moments has been developed[18] for light nuclei (mass number between 6 and 16), using the individual particle model (Hartree model). By assigning quantum states and individual wave-functions to each nuclear particle, one can calculate, in the same manner as for atomic spectra, the energy of the various terms which arise from any given configuration of neutrons and protons. The important feature for nuclear moments is to predict which of the spectroscopic states of the nucleus will be the most stable. The nuclear magnetic moment can then be calculated from considerations as to the coupling of the spin and orbital momenta of the nucleus in the ground state. The calculations involve the same specific nuclear forces as are used in other phases of nuclear physics, and assume that the magnetic moments of the proton and neutron are not modified by the nuclear binding process. Thus the moments are predicted without any empirical adjustment of parameters. The observed I values, including $I=\frac{1}{2}$ of ^{15}N determined recently from the band spectrum[19], agree with those predicted. The agreement for the magnetic moment is fairly satisfactory in the individual cases except ^{14}N, but the predicted ratios and differences of magnetic moments of those nuclei for which accurate direct determinations are available do not agree with the observed values within the experimental error[15].

REFERENCES

(1) *Reports on Progress in Physics, Phys. Soc.*, **1**, 110, 306 (1934); **2**, 81, 346 (1935); **3**, 71, 371 (1936).
(2) Jackson, D. A. and Kuhn, H. *Proc. Roy. Soc.* A, **165**, 303 (1938); and also references quoted here.
(3) Meissner, K. W. and Luft, K. F. *Ann. Phys., Lpz.*, **28**, 667 (1937); Meissner, K. W. *Ann. Phys., Lpz.*, **31**, 505 (1938).
(4) Jackson, D. A. and Kuhn, H. *Proc. Roy. Soc.* A, **164**, 48 (1938).
(5) Heyden, M. and Ritschl, R. *Z. Phys.* **108**, 739 (1938).
(6) Ritschl, R. and Schober, H. *Phys. Z.* **38**, 6 (1937).
(7) Kopfermann, H. and Krüger, H. *Z. Phys.* **105**, 389 (1937).
(8) Heyden, M. and Kopfermann, H. *Z. Phys.* **108**, 232 (1938).
(9) Millman, S., Rabi, I. I. and Zacharias, J. R. *Phys. Rev.* **53**, 384 (1938).

(10) RABI, I. I., ZACHARIAS, J. R., MILLMAN, S. and KUSCH, P. *Phys. Rev.* **53**, 318, 495 (1938).
(11) BETHE, H. A. and BACHER, R. *Rev. Mod. Phys.* **8**, 82 (1936). [Summarizes the work in this field.]
(12) BREIT, G. *Phys. Rev.* **51**, 248 (1937).
YOUNG, L. A. *Phys. Rev.* **52**, 138 (1937).
(13) BREIT, G. and RABI, I. I. *Phys. Rev.* **38**, 2082 (1931).
(14) RABI, I. I. *Phys. Rev.* **51**, 652 (1937).
(15) BETHE, H. A. *Phys. Rev.* **53**, 842 (1938).
(16) SCHMIDT, T. *Z. Phys.* **108**, 408 (1938).
(17) KORSCHING, H. *Z. Phys.* **109**, 349 (1938).
(18) FENNBERG, E. and WIGNER, E. *Phys. Rev.* **51**, 95 (1937);
ROSE, M. E. and BETHE, H. A. *Phys. Rev.* **51**, 205 (1937);
INGLIS, D. R. *Phys. Rev.* **53**, 470, 880 (1938).
(19) KRÜGER, H. *Naturwissenschaften*, **27**, 445 (1938).

III. THE STARK EFFECT AND SOME RELATED PHENOMENA

By J. STUART FOSTER, F.R.S., Macdonald Professor of Physics, McGill University, Montreal

§ 1. INTRODUCTION

The simple atomic model used in the classical theory of the normal Zeeman effect denied an electric analogue. Voigt (1901) nevertheless interpreted an electric double refraction reported by Kerr (1875) as evidence for small changes in the electronic periods. These changes were attributed to a slight modification in the law of force on the electron when the equilibrium position of the latter is greatly displaced by an external electric field. Only one feature of this elementary theory of the electric effect has persisted through modern developments, viz. the asymmetric splitting and displacements of spectral lines. The theoretical viewpoint was indeed completely changed by the appearance of the Bohr frequency relation in June 1913. This relation offered the first encouraging basis for a search for relatively large electric displacements.

The discovery of the electric effect in a modified canal-ray tube was announced in November 1913 by Stark, who already had made observations on both hydrogen and helium atomic spectra in fields of 13 and 31 kv./cm. The components of H_β and H_γ were found to be *polarized* and apparently were displaced *symmetrically* with respect to the original line by amounts *proportional to the external electric* field. None of these features was found in the helium spectrum, where the *magnitude* of the displacements was seen to depend to some extent upon *series relationships* while the *relative intensities* of "components" varied with the field. An independent announcement by Lo Surdo (1913) confirmed Stark's observations on the H spectrum, and revealed the experimental importance of the cathode fall in small tubes for the production of electric fields.

The order of magnitude of the displacements of the outer π components in H as well as their polarization received an immediate theoretical interpretation by Bohr, who also gave a qualitative explanation of the smaller effects observed in He.

§ 2. FINE ANALYSIS OF THE STARK EFFECT IN ATOMIC HYDROGEN AND THE QUANTUM THEORY INTERPRETATION

Weak traces of additional components in his "rough" analysis led Stark (1914) to re-examine the Balmer lines from a strong source in electric fields of 74,000 v./cm. and 104,000 v./cm. Light from an improved low pressure canal-ray tube was analysed by very fast plane grating and prism spectrographs of moderate dispersion. A surprisingly large number of polarized components was found in symmetrical patterns with a range of relative intensities so great that faithful reproduction of complete analyses is extremely difficult.

An explanation of the linear displacements was provided by Epstein (1916) in an extension of the Hamilton-Jacobi method. Using parabolic co-ordinates together with the angle measured round the field axis, the variables are separable under Newtonian mechanics though the fine structure of the original lines is thereby lost. An original level n is split symmetrically into $2n+1$ Epstein levels with the common separation

$$n\frac{3hF}{8\pi^2mec}\ \text{cm}^{-1}$$

between adjacent members of the group.

The selection rules and polarizations of the hydrogen components were placed on a reasonable basis by Bohr in an application of perturbation theory and the correspondence principle. It is learned that the time mean position of the electron moves in an ellipse perpendicular to the field axis with a frequency

$$\omega_F = n\frac{3hF}{8\pi^2mec}\quad\left(\text{cf. }\omega_H=\frac{H}{4\pi mc}\right).$$

The frequency of revolution, ω, and ω_F are introduced into the expression

$$\delta E=\omega\delta J+\omega_F\delta J_F$$

to give the energy found by first approximation in the Epstein theory and to show that the planes in which the electrical centre moves cut the field axis at points with spacing $\frac{3}{2n}$ (major axis of ellipse). The above analysis of the motion is interpreted through the correspondence principle to give the selection rules and polarizations.

Rather good estimates of the relative intensities of Stark components in H were based by Kramers (1919) on his quantitative harmonic analysis of the motion of the electron. Moreover, in an application of perturbation theory to include relativistic corrections, Kramers (1920) showed the connexion between observed Stark components and the fine structure of H_α.

§ 3. EXPERIMENTAL DEVELOPMENTS AND QUALITATIVE INTERPRETATIONS OF RESULTS

Although the facts and the interpretation of the Stark effect in H were apparently known soon after its discovery, a delay in reaching an understanding of the effect in other spectra was caused by a necessary development of Lo Surdo sources and by the gradual evolution of theoretical ideas.

Fields comparable to those realized in canal-ray sources were established by Lo Surdo through the application of a few thousand volts to a simple two-electrode tube 1·5 mm. in diameter at the cathode. From such a source Brunetti[1] first observed "isolated components" in He—now known as entirely separate electric combination lines.

Stark and Kirschbaum met with some success in their analysis of the effect in metallic spectra which they excited by placing chlorides on the auxiliary electrode of a canal-ray tube containing oxygen. Nevertheless the apparently progressive "splitting" reported for successive members of diffuse series serves to illustrate the difficulty of detecting with this source the origin of the electric combination lines which actually lie close to the original diffuse lines and are entirely responsible for the apparent "splitting". This difficulty was not encountered by Koch in his recognition of the well-separated He lines $2\,^3P - n\,^3P$.

An impression that the electric effect might be of quite different character in the canal-ray and in Lo Surdo sources was dispelled by Nyquist[2] who employed a tube of modified Lo Surdo type and made important detailed observations on the visible spectra of H, He and Ne. A symmetrical splitting was observed for the He^+ line $\lambda 4686$.

Following Nyquist, Stark and his associates carried the investigation into the ultra-violet. Many of the He electric combination lines for this region were observed and given special names. Finally, the Copenhagen analysis of the Stark effect in Hg[3] gave a clear indication of the general interpretation of the new combination lines on the Bohr theory (Δk = any value in an electric field of suitable strength).

Owing to the linear Stark effect in H, the outer components of the Balmer lines may be used to measure the distributions of the electric field strengths in the bright beam which traverses the Crookes dark space in Lo Surdo sources. Starting from the end of the dark space and going toward the cathode, the field varies directly with the distance in a large tube, and as the square of the distance in quite small tubes used by Takamine and Yoshida. Actually the function is commonly more complex, and may be to some extent controlled by the exact size, shape and spacing of parts near the cathode in special modified Lo Surdo sources.

In hydrogen canal-ray tubes having the electric field parallel to the motion of the protons the outer π components toward the red were found to be stronger than their mates (Stark-Lunelund effect). The asymmetry reversed with the field, and was explained by Bohr on the basis of relative losses of initial Epstein states in collisions.

The second-order red-shift for hydrogen components as predicted in the Epstein theory by a second approximation is proportional to the square of the applied field and the sixth power of the principle quantum number with relatively small variations among components of a Balmer line. This shift was detected by Takamine and Kokubu[4] for the central component of H_γ on an early Lo Surdo plate taken at 147 kv./cm.

Much general information about the Stark effect in metallic spectra was obtained through the use of a bell-jar model of the Lo Surdo source as developed by

Anderson[5] for an investigation of the Cr spectrum. The relatively large cathode of the metal to be studied was surrounded by quartz and run red hot. Using cathodes of moderate size, Takamine[6] obtained fields as high as 75 kv./cm. Among the points brought out in these studies is the fact that early members of a new electric combination series are commonly displaced toward the violet while later members are gradually shifted to the red.

§ 4. PATTERNS OF THE STARK EFFECT

It was noticed by Foster[7] that Kramers' theory of the connexions between the original Sommerfeld fine structure and the Epstein components implied that the members of a fine-structure "series" should present a constant pattern in respect to numbers of components and their polarizations. Since the line structure of parhelium (singlet system) is sufficiently open to allow observations on the splitting of individual lines, the He spectrum was examined for patterns. The Lo Surdo source designed for this work gave sharp components at all fields from zero to the maximum value. Analyses were carried out with a six-prism glass spectrograph of high dispersion. Since the known Stark effect in H appeared on all plates, this was used incidentally to determine field strengths, and thereby to plot the very complex relations between field strength and displacement which exist within even a single group of He lines at moderate fields. Indeed such relations, once established for initial He levels, were used to identify beyond question certain electric combination lines which appeared only in high fields and with rather large displacements. This research established the normal Stark patterns[8] which are briefly described as in accord with modern selection rules, it being understood that levels $\pm m$ are coincident. This result, compared with the theory of Kramers, at the same time established the fact that the Stark effect in He is not dissimilar from that in H as previously supposed, and reacted to emphasize the need of relativity corrections for an understanding to the Stark effect in H.

The theory of the Stark effect in helium as carried out by Foster[9] is a typical example of the perturbation theory in the quantum mechanics of Heisenberg. In this theory the H atom is perturbed by two forces (i) a central field which is asked to account for the difference between He and H terms identified by the same quantum numbers (n, k, m) as given by tables, photographs, or calculated, and (ii) the external field which contributes terms in the energy matrix exactly as in the case of the Stark effect for the H spectrum. From general theoretical considerations (the selection rule for m) it is clear that only terms having the same m value will repel one another in the electric field. For each group of terms with fixed n, m the theory leads to an equation of the $(n-m)$th degree whose roots are the displacements of the $(n-m)$ He terms contained in the group. When the displacements are measured from a normal He line it is revealed that the *separations* of He terms are the important factors determining Stark effects rather than the absolute displacements of the terms from the corresponding terms of H.

It follows that, in contrast with the Zeeman effect, the term with highest m value within the group is not displaced, whereas the outer terms of the group with $m = 0$ are most displaced. The situation in general is rendered complex by the more

or less random separations of the He terms and the fact that the displacement of a given member will depend largely upon its position with reference to others in the group. Nevertheless the observed displacements and intensities are in good agreement with the theory.

In very high fields where the displacements are large compared with the initial separation of the terms under consideration, it is learned that the displacements should follow a linear relationship with the field while the relative intensities should settle down to steady values after their fluctuations in moderate fields. Photographs show that this condition may be closely approached in the laboratory.

§ 5. STARK EFFECT IN RARE GASES

Investigations of the Stark effect in other rare gases have been carried out in Montreal by Foster and Rowles[10] (Ne), Harkness and Heard[11] (Xe) and Foster and Horton[12] (Kr and A). In a brief general review[12] of the effect in all these spectra energy-level diagrams are included to represent graphically how the atomic levels repel one another in the field. It is noted that in the Ne spectrum terms with different limits cross each other in the field with no sign of repulsions, whereas in A some displaced terms are turned back upon the close approach of terms with the other limit. This is attributed to the stronger probability of transitions between the two sets of terms in A. It is not uncommon to find lines displaced in one direction at low fields and eventually pushed back to the other side of the original line at higher fields.

The patterns of the Stark effect in all these spectra are those determined by the j values of initial and final states with a few possible exceptions. The majority of the patterns are not fully resolved. Similar patterns were found by Haslam[13] who re-examined the Hg spectrum in fields of sufficient strength to resolve Stark components.

§ 6. FINER DETAILS OF STARK EFFECT IN ATOMIC HYDROGEN

While the fine structure of the H lines does not appear on Lo Surdo photographs, McRae[14] was nevertheless able to demonstrate very clearly that certain fine-structure separations persist in high fields and cause variations from the Epstein spacings which are readily detected by the eye alone. This completes the experimental evidence which forces one to consider the fine structure as a vital factor in the Stark effect for H.

The redrafting of the Epstein theory by Schrödinger[15] gave quantitative relative intensities which were only in rough agreement with the experimental values by Stark. In contrast to this, the new values were at once seen to be in entirely satisfactory agreement with results of a Lo Surdo analysis by Foster and Chalk[16]. Nevertheless Mark and Wierl[17] demonstrated beyond question that the Schrödinger intensities are not in general found in canal-ray tubes, and indeed that the actual intensities vary with the relative directions of canal rays and applied field as well as with the direction of observation. The results are only partially understood on the basis of selective depletion of initial states in collisions.

Clear indications of a gradual change from canal-ray characteristic intensities

toward theoretical values were obtained by Thornton[18] with a canal-ray tube run at pressures up to 0·5 mm., or about twenty times the usual value. Finally Foster and Snell[19] found an asymmetric D_α which appeared as the mirror image of an asymmetric H_α in the same Lo Surdo photograph; this interesting result appears to be due to self-absorption.

In conclusion one may say that the Schrödinger intensities are best realized by means of a Lo Surdo source containing helium (or other rare gas) with only a trace of hydrogen, and that the great variations from theoretical values which undoubtedly exist are mainly due to physical conditions not considered in the theory.

§ 7. STARK EFFECT COMPARED WITH POLE EFFECT IN IRON

On the basis of very little early evidence it was for a long time believed that the shift of Fe lines noticed in light taken from near the pole of an iron arc (pole effect) was in reality only a Stark effect. Following a recent analysis of the Stark effect in Fe by Panter and Foster[20], this view is left without any support whatever. No quantitative or even qualitative connexion is established between the pole effect and the effect due to a uniform external electric field; indeed, for a few lines the displacements are in opposite directions. The observed Stark effect in Fe is limited to even terms whereas the pole effect is not so limited. A noteworthy feature in this investigation is the source, so modified from the Anderson type by reduction of cathode diameter as to take only 2 mA. current and produce fields upwards of 150,000 v./cm.

§ 8. STARK EFFECT IN *B*-TYPE STARS

Struve[21] first reported the presence of an absorption line at λ4470 in the spectra of certain *B*-type stars, and this he immediately identified with the helium electric combination line $2\,^3P-4\,^3F$. After a quantitative comparison of this line with $2\,^3P-4\,^3D$ (λ4471) at the Dominion Astrophysical Observatory, Victoria, Foster and Douglas[22] concluded that the Stark effect offered no explanation of the great strength observed at the zero field position of $2\,^3P-4\,^3F$. In a complete report of this investigation to be published shortly it is suggested that the "zero-field" absorption should not be attributed to $2\,^3P-4\,^3F$, but rather be considered a minimum in the contour of a very broad $2\,^3P-4\,^3D$ line left after emission has worn away the line in irregular fashion through resonance between $4\,^3D$ and $4\,^3F$. Actually the absorption line $2\,^3P-3\,^3D$ (λ5876) is weak. Evidence for Stark effects is nevertheless found at high fields and in displacement of $2\,^3P-n\,^3S$ lines.

§ 9. STARK EFFECT IN MOLECULAR SPECTRA

Since the earliest experiments it has been known that the Stark effect in the spectrum of molecular hydrogen is small and irregular. A rather extended early research was published by Kiuti[23]. Somewhat broader investigations of the molecular spectrum have been carried out in Montreal by Snell[24] (H_2), Jones (HD, unpublished) and Neamtan (D_2, unpublished). The theory has been given by Macdonald who finds that the Stark effect depends upon a second-order per-

turbation calculated through the use of suitably selected wave functions. Detailed numerical comparisons with the experiments have not been made.

Streaming electrolytic hydrogen through a McGill Lo Surdo source and making the analysis with a 30 foot stigmatic grating spectrograph, Snell found 350 displaced lines between H_α and H_β. Several of these showed the full quota of components which are expected if the total angular momentum K is resolved to give $K+1$ electric levels ($M=0, \pm 1, \pm 2 \ldots \pm K$; $\Delta M=0, \pm 1$) with appreciable separations realized only in the initial states. Nevertheless incomplete patterns are more commonly found, and this fact prevents a complete deduction of the quantum numbers associated with each observed component. There are, of course, many cases of lines associated with the same initial state which consequently show like displacements of some components. A rather general large blue-shift in the system $3d\,^3\Delta - 2p\,^3\Pi$ indicates an electric perturbation in $3d\,^3\Delta$.

Neamtan has examined about 175 displaced lines in the D_3 spectrum and of these 79 are known to be Fulcher-band lines. As in atomic spectra, the theory claims that levels which lie close together and between which transitions are possible will repel one another in electric fields. On this basis, and in view of the energy-level diagram as known to date, there is reason to expect that the levels of the $2s\,^3\Sigma$ state may be elevated while those of the $3p\,^3\Pi$ state are depressed. The Fulcher-band lines which originate from transitions between such states may therefore be expected to show displacements toward the red, and generally speaking this is actually the case.

In the HD spectrum Jones examined about 125 lines of which 70 are in Fulcher bands, with the displacements predominantly toward the red. There is no clear evidence of a linear effect in this spectrum.

In all of the above researches there appears to be good evidence that the final states are sometimes affected by the field.

§ 10. EFFECTS DUE TO COMBINED ELECTRIC AND MAGNETIC FIELDS

The technique of obtaining analyses of spectra when the source is subjected to combined electric and magnetic fields is similar to that required in studies of the Stark effect alone, since in each case a possible instability of the source offers the chief difficulty. This possibility arises from the fact that ion density and distribution may determine the electric fields wholly or in part, whereas such factors have no appreciable effect on the magnetic field. Nevertheless when the fields are parallel Foster[25] found that a Lo Surdo source threaded through the hollow poles of an electromagnet was uncommonly stable and that the analyses obtainable were in every way equal to those found in pure Stark effect.

In parhelium the magnetic separation (in a field of 15,000 gauss) of σ components is exactly the normal Zeeman separation throughout the wide range of electric fields whereas the π components are unaffected by the magnet. In accord with theory, and within experimental error, the two effects are strictly additive. Parallel fields offer a ready means of experimental proof that the components in most normal Zeeman effects are complex.

The effects are not exactly additive in orthohelium. Some of the Stark components previously observed for the weakest member of the fine structure of diffuse-series lines are suppressed by the magnet and the exact character of the one remaining "σ" component is determined. Moreover, the two components into which strong σ components are split are certainly not of equal intensity, as in parhelium. In equally clear analyses for neon recently carried out in Montreal by Prof. W. Rowles and Dr L. Rowles the departures from the additive condition are till more marked.

For the study of effects due to electric and magnetic fields crossed at right angles Foster employed a Lo Surdo tube with coaxial cylindrical electrodes set with axis parallel to the magnetic field. By means of suitable insulation the radial discharge was allowed to strike only a very short section of the central cathode. This design permitted the use of cathodes large enough for effective cooling, and at the same time gave a range of electric fields from 0 to 10,000 v./cm. Among the features in helium are (i) no observable splitting of the $2\,^1P-n\,^1P$ series, (ii) very wide magnetic splitting of the series $2\,^1P-n\,^1F$ and (iii) rapid changes in relative intensities of components with increasing electric field. Many structures were obviously very complex and unresolved. The observations were in the direction of the magnetic field. With a somewhat stronger source the research was extended by Foster and McRae. Some of the improved analyses have been reproduced by Brunetti[1]. In the process of this extension a special canal-ray tube was developed in which the canal rays ran along the axis of the magnetic field from each direction while between the poles an electrostatic field was applied at right angles. The direction of observation was perpendicular to each field. Faint analyses obtained with this source were characterized by a rather common lack of symmetry in the intensities of the magnetic components.

In an unpublished McGill Thesis by Pounder this work has received a further marked extension by the same general method. The poles themselves being used as cathodes, the magnetic field has been increased while the electric field ranges up to 120 kv./cm. This is sufficient to separate many Stark components associated with transitions from initial states $m = \pm 2$, and in some cases, e.g. a satisfactory magnetic splitting, shows three components, or in other words reveals a transition in which $\Delta m = \pm 2$. This type of transition is not, however, perfectly general, since no components corresponding to $m = \pm 3 \rightarrow \pm 1$ are observed, though such components if present could easily be detected on the well-exposed plates. A complete report of this most detailed analysis in crossed fields will shortly be published.

It is expected that the large cyclotron magnet now under construction at McGill University will be used temporarily for further researches on effects of combined electric and magnetic fields.

REFERENCES

(1) Brunetti, R. *L'Atome e le sue Radiazioni* (1932).
(2) Nyquist, H. *Phys. Rev.* **10**, 226 (1917).
(3) Hansen, Takamine and Werner. *K. danske vidensk. Selsk. Math.-fys.* **5**, 3 (1919).
(4) Takamine, R. and Kokubu. *Mem. Coll. Sci. Kyoto*, **3**, 275 (1919).
(5) Anderson, J. A. *Astrophys. J.* **46**, 104 (1917).

(6) TAKAMINE, R. *Astrophys. J.* **50**, 23 (1919).
(7) FOSTER, J. S. *Phys. Rev.* **23**, 667 (1924).
(8) FOSTER, J. S. *Proc. Roy. Soc.* A, **114**, 47 (1927).
(9) FOSTER, J. S. *Proc. Roy. Soc.* A, **117**, 137 (1927).
(10) FOSTER, J. S. and ROWLES, W. *Proc. Roy. Soc.* A, **123**, 80 (1929).
(11) HARKNESS, H. and HEARD, J. F. *Proc. Roy. Soc.* A, **139**, 416 (1933).
(12) FOSTER, J. S. and HORTON, C. A. *Philos. Trans.* **236**, 473 (1938).
(13) HASLAM, R. N. H. *Proc. Roy. Soc.* A, **150**, 338 (1935).
(14) MCRAE, D. R. *Proc. Roy. Soc.* A, **132**, 257 (1931).
(15) SCHRÖDINGER, E. *Ann. Phys., Lpz.*, **80**, 437 (1926).
(16) FOSTER, J. S. and CHALK, L. *Proc. Roy. Soc.* A, **123**, 108 (1929).
(17) MARK, H. and WIERL, R. *Z. Phys.* **53**, 526 and **55**, 156 (1929).
(18) THORNTON, R. L. *Proc. Roy. Soc.* A, **150**, 259 (1935).
(19) FOSTER, J. S. and SNELL, A. H. *Proc. Roy. Soc.* A, **162**, 349 (1937).
(20) PANTER, S. F. and FOSTER, J. S. *Proc. Roy. Soc.* A, **162**, 336 (1937).
(21) STRUVE, O. *Observatory*, **61**, 53 (1938).
(22) FOSTER, J. S. and DOUGLAS, A. V. *Nature, Lond.*, **134**, 417 (1934).
(23) KIUTI. *Jap. J. Phys.* **4**, 13 (1925).
(24) SNELL, A. H. *Philos. Trans.* **234**, 115 (1935).
(25) FOSTER, J. S. *Proc. Roy. Soc.* A, **131**, 133 (1929).

IV. ASTRONOMICAL SPECTROSCOPY

By A. HUNTER, Ph.D., F.R.A.S.
Royal Observatory, Greenwich

§1. CORONAL LINES

Another page in the chequered history of "coronium" has been turned in the last year by the discounting of yet another theory concerning the origin of the emission lines in the spectrum of the solar corona. We are still as ignorant as ever of the nature of the emitting particles.

For only a few years after the discovery in 1869 of a green emission line in the corona did the theory that it was identical with the chromospheric line at 5317 A. pass muster. Early objections based on the improbability of this singling out of a particular iron line were supported as soon as accurate wave-length measurements at the turn of the century put the wave-length of the coronal line at 5303 A. At practically all total eclipses since then, at least one expedition has included in its programme the examination of the coronal spectrum. Some 20 lines are now known, not all of which, however, appear on all occasions. This latter fact suggests a real difference of relative intensity at different epochs which may be correlated with other temporal changes of the corona, such as its change of shape during the sunspot cycle. Hopelessly entangled with this effect are instrumental uncertainties which make the lot of the spectroscopist who attempts to recognize multiplet relationships from the results an unhappy one. The wave-length measurements no doubt serve his purpose well enough now, but most of the published intensities depend on eye estimates which do not take into account such factors as the spectral sensitivity curve of the photographic emulsion used, or the dispersion of the optical system employed. Standardization of the intensity scale is an urgent need which may prove a *sine qua non* for future progress.

The green line is so much stronger than any of the others that attention has naturally been concentrated on obtaining its wave-length with great accuracy. At best, of course, it is intrinsically faint, and the time available for recording it short, but nevertheless the best observations agree to within ±0·10 A. on 5302·92 A. Interferometric observations have met with little success, partly owing to the intensity difficulty, but mostly because the great width of the line necessitates the use of small separations between the interferometer plates. Observations out of eclipse have the advantage that time allows the taking of many spectrograms and thus reducing the probable error of the result, but here again the practical limit is set by the intrinsic width of the lines. Lyot, who minimizes scattered light in the atmosphere by observing at an altitude of 10,000 ft., and in his instrument by taking immense pains in the design of his optical train, has so far the monopoly of this kind of observation. His most recent result for the green line is 5302·83 ± 0·02 A. The other lines are not so well known and their wave-lengths can hardly be regarded as certain to a quarter of an angstrom.

The problem of finding the emitter is helped by slitless observations. Monochromatic images of the corona at 5303 and 6374 A. show structural details resembling the eruptive prominences, and suggest that although the inner corona is the main seat of the emission lines, extensions to as much as a solar diameter may occur at certain points, notably those from which, in the image in integrated light, there appear the characteristic rays looking rather like the lines of force round the pole of a bar-magnet.

The identification problem has been attacked from two main directions. One method—the older—depends on finding coincidences with known spectra. The trouble with this procedure is, of course, that some chance coincidences in the form of observed lines, and many in the form of intercombinations, likely or otherwise, between known term values, can usually be found by the assiduous searcher; the more so if a large tolerance is permitted in order to cover experimental error in the measured wave-lengths. This method of approach has led to some premature claims: the lines have been attributed in turn to iron, to "proto-fluorine", to singly- and doubly-ionized calcium, to argon, and to oxygen, neutral and ionized; all with a certain amount of supporting evidence, and the last three within the past decade. None of these assignments has received general support, however, mainly because no one theory fits more than a few of the observed facts.

The second method has more to recommend it. Certain elements, notably H, He and Ca, are known to exist in the high chromosphere. Certain others might reasonably be expected there. The physical conditions, as to temperature, radiation flux, pressure, etc., are fairly well known for this region, and can be extrapolated to coronal distances without, one can hope, introducing serious errors. Thence the electron transitions which might be expected in these elements under these conditions can be predicted, and if the terms of the atom or ion concerned are well enough known, approximate wave-lengths can be derived. It was this line of approach which led Bowen to his conclusion, since abundantly justified, that certain of the lines in the galactic nebulae are due to forbidden inward transitions

from metastable states of doubly-ionized oxygen and nitrogen. The pressure in the corona must be extremely small, so that from this point of view, at any rate, atoms might be expected to persist long enough in metastable states for forbidden transitions to occur notwithstanding their low transition probabilities. So far, however, the search has proved unsuccessful (though of course even negative results have their value): a survey of forbidden lines in Fe II, Ca II, Ti I, Ti II, Ti III and A II shows no certain coincidences. The trouble here is probably that the radiative flux is too high in the corona for metastability to continue long enough for the production of forbidden lines; the atoms concerned are too likely to absorb quanta before emitting. Once they have reached stable states by this process, allowed transitions will take place and the resulting lines will presumably be indistinguishable from chromospheric lines.

Until a year or so ago, perhaps the most hopeful suggestion[1] was based on the great width of the lines, which is not a feature of the nebular lines. Line-widths of the order of angstroms suggest that the excited states involved have unusually short life-times—quite the reverse of metastability. It may be, then, that we are dealing with double-electron transitions from an excited state in which two electrons are simultaneously removed to outer orbits.* Approximate calculations[2] made along these lines for helium, the most probable element concerned, show qualitatively that diffuse lines may be expected in the visible region, though of course no exact wave-lengths can be predicted. Support is lent to the theory by the recent discovery that helium and "coronium" co-exist in the upper chromosphere. The absence of ordinary He and He^+ lines from the coronal spectrum, however, awaits satisfactory explanation, as does also the fact that no Rydberg series can be found from the measured wave-lengths.

In the absence of more supporting evidence, it will be seen that this theory stands or falls by the explanation offered for the contours of the coronal lines. In a recent paper[3] Waldmeier accounts for their great breadths in a totally different way. He points out that the equivalent widths of the stronger lines are approximately proportional to their wave-lengths. May not the broadening, then, be a Doppler effect? Mere thermal broadening, by Brownian motion of the particles concerned, cannot be the only agent, although the line-contours would support such an hypothesis, because the temperature needed is impossibly high. There remains the possibility of an outward radial streaming of the atoms. Unfortunately the contours due to this cause alone do not fit at all well with those observed, and neither thermal broadening nor rotational broadening, nor both combined, can modify the curve enough to make the two contours fit. If the stream velocity is not homogeneous, however, the outward motion being a combination of several velocities, the observed curve can be reproduced exactly.

The theory is not without its knotty points, of course, and more observational material, particularly with regard to the profiles of the infra-red lines, is needed to test it. It has the great advantage, however, that the mechanism it demands is not wholly arbitrary. We know that solar agencies exist capable of producing eruptive

* See also p. 218.

prominences, which move at speeds of the order required (50 km./sec.), and may on occasion[4] reach points more than a solar diameter away from the limb, i.e. well beyond that portion of the corona responsible for line emission. We know also that monochromatic images of the corona in the light of its own green and red lines show structural features closely resembling eruptive prominences. And, finally, the Fraunhofer lines in the corona exhibit displacements which can be explained[5] by assuming an outward velocity of the scattering particles of the order of 30 km./sec. It is independently probable, then, that some such streaming does occur in the corona. If this is indeed so, we need look no further for an explanation of the exceptional width of the coronal lines; the double-electron transition theory loses its main support; and the whole question of the origin of the coronal lines is back in the melting-pot.

§2. REDDENING IN *B*-TYPE STARS

It has been known for some 15 years that the colour temperatures deduced from measurement of the spectral distribution curves of *B*-type stars are by no means as homogeneous as would be expected for members of a single spectral class. The early *B* stars seem to be responsible for most of the scatter. Normal *B*2 stars, for instance, have colour temperatures of over 25,000° K., whereas such a star as χ^2 Orionis, though in the same spectral subdivision, seems to be at only 8000° K. The distribution in the continuous spectra of certain other fainter *B*-type stars resembles that in stars of types as late as *G*5. This enormous dispersion in colour temperature implies a challenge which the astrophysicist has not been long in taking up.

The reason for the reddening is by no means certain yet. There is one point, however, which should be made clear at the outset. The measurements of colour temperature are measurements essentially of colour, the word "temperature" being merely a convenient and well-recognized means of expressing the results. Most astronomers would in fact be reluctant to admit that such low temperatures could in any way represent the real conditions prevailing in stellar atmospheres which are capable of absorbing *B*-type spectra. We are therefore concerned with finding a cause not so much for the anomalous temperatures of certain *B* stars as for their anomalous colours.

There exists a considerable body of evidence pointing to the occurrence in interstellar space of matter capable of absorbing or scattering visible radiation. The dark nebulae which obscure the background of faint stars in certain regions of the sky are evidence of one kind of absorption—a non-selective blocking process probably due to gross particles of meteoric type; the "detached" lines in the spectra of the more distant stars are now attributed to another—the absorption of discrete wave-lengths by atoms, neutral and ionized, and perhaps also by molecules. It is tempting to conclude that the anomalous colour of the "yellow" *B* stars is due to selective absorption by particles intermediate in size between these two. If such particles were small in relation to the wave-length of visible light, they would produce Rayleigh scattering proportional to λ^{-4}, which corresponds, at least in a

rough qualitative way, with that observed. There are *a priori* grounds for ascribing the reddening to scattering by interstellar matter. The *B* stars, being of high luminosities, are on the average at a great distance, and would be expected to show any space absorption which occurs. And indeed stars of the greatest colour excess do tend to be found in the galactic plane, where such absorption would presumably be at a maximum.

Experimental support for the hypothesis came from Greenwich in 1928, when it was pointed out that the abnormal colour temperatures were correlated with the intensities of the interstellar *K* line, strong detached lines going with low temperatures. On this basis, then, the effect is attributed to selective absorption by material coextensive with the cloud of calcium ions which is responsible for the ultra-violet interstellar lines. Two years later, the work of Trumpler on open clusters, though not primarily connected with *B* stars, shed further light on the matter. Estimating the distances of the clusters from their apparent diameters, he found a progressive increase of colour excess in the component stars with increasing distance. There can be no doubt, then, that space reddening does occur in the galactic plane. But this is, of course, no proof that the reddening in *B* stars is caused in this way. The most recent work on the subject comes from Harvard[6], where Greenstein has made a spectrophotometric study of 38 reddened *B* stars. The magnitude differences between these stars and normal *B* stars along the spectrum from 6500 to 3500 A. indicate that the law of interstellar absorption is such as to make the apparent energy distribution of a reddened star very similar to that of a black body at a lower colour temperature. What little departure there is from this rule makes the colour temperature rise somewhat towards the violet; in no case does there exist the progressive decrease which would be expected for strongly selective absorption. The absorption is thus certainly not of the Rayleigh type: it is more nearly linear with frequency. As regards its cause, Greenstein finds from a study of the known regions of obscuration in the sky that obvious dark clouds surround every one of the 38 stars on his observing list. The temptation to ascribe the effect to these clouds is irresistible. In a theoretical interpretation[7] of the observed results, he suggests that the clouds are composed of mixtures of metallic and dielectric particles of various sizes, the high selectivity of the extinction by small particles being reduced by the general scattering by larger particles. A particularly attractive result of this theory is that the observed selectivity is given when the frequency function of the particle sizes is an extrapolation of that observed for hyperbolic meteors.

If this were all the available evidence bearing on the subject, we could have little doubt as to the cause of the reddening. Unfortunately, however, only one side of the question has yet been touched upon. The first doubts were voiced by Gerasimovič, who pointed out that the reddening in some of the yellow *B* stars was associated with diminished absorption starting sharply at the Balmer limit and extending into the ultra-violet. May not the reddening, then, be inherent in the stars concerned? Such a view is partially confirmed by the results of Williams[8], who examined quantitatively the relationship between colour excess and detached-line intensity in *B*-type stars. The correlation goes in the way to be expected if

space absorption causes the reddening: large colour excesses are associated with strong interstellar lines. Williams's work, however, emphasized a distinction of which the importance is being realized more and more as fresh data accumulate: the *B* stars showing, as is normal, absorption lines, give smaller colour excesses than do the *Be* stars, which show emission lines. In fact, the stronger the hydrogen emission, the greater the residual colour excess, i.e. the lower the colour temperature as compared with normal *B* stars at the same distance. The implication is that distant stars showing *B*-type absorption are space-reddened; those showing *B*-type emission are similarly space-reddened, but are in addition intrinsically yellow to an extent depending on the strength of their emission lines.

This idea of a dual mechanism for the reddening has been supported quite recently by the peculiar behaviour of the slightly reddened *Boe* star γ Cassiopeiae. This star has long been suspected of irregular spectrum variability, but the middle of 1936 saw the beginning of altogether unprecedented variations[9], both in the line emission spectrum and in the background continuum. The star brightened by half a magnitude, its colour temperature fell from the normal 16,100 to 10,500° K. at the end of 1936, and further to 9200° K. during 1937, and the intensity of the hydrogen emission lines increased, that of Hγ going up five-fold by the middle of 1937[10]. We have evidently here been given the opportunity of actually watching the process of reddening being carried a stage further. Obviously in this case the increase cannot be a distance effect: the radial velocity of the star would have to be incredibly big. Nor can we put it down to the intercalation of a relatively dense cloud between us and the star, for the star has brightened. Some physical mechanism associated with the star itself must be postulated to increase the intensity of the emission lines and redden the continuous spectrum simultaneously. Such a mechanism might also, by the way, serve to explain why Nova Herculis in its early stages had an early *B*-type emission spectrum, but an *A*-type colour temperature[11]. It may be that the *Be* stars are surrounded by extended chromospheres in which selective scattering takes place. We have independent evidence for the expulsion of such atmospheres from novae. Woolley, however, has pointed out that the effect may be as well explained by the addition of red light as by the absorption of blue. His suggestion[12] is that fluorescent emission occurs in the stellar atmosphere throughout the visible spectrum, coming to a maximum at the Paschen limit. On this view, fluorescent excitation of the visible Balmer lines would also be expected; the observed increase in intensity of Hγ may well be due to this cause.

The exact nature of the mechanism postulated, however, does not alter our main deductions from the various observations: the abnormally low colour temperatures of certain *B* stars are due partly to selective absorption in interstellar space, and partly to a reddening effect, inherent in the stars concerned, which seems to be associated with line emission.

§3. THE SUPERGIANT ϵ AURIGAE

The variable star ϵ Aur has been an astronomical enigma for over a century. It is an eclipsing binary of such a long period that its exact nature remained in doubt until its latest minimum in 1929–30—the fourth since discovery of its variability.

No fewer than 13,800 photometric observations made by 47 observers since 1842 are now available[13], and establish a period of 27·1 years. The periods of most eclipsing variables range from a few hours to a few hundred days; a value of nearly 10,000 days is unique. For about 25 years the visual magnitude of the star remains steady at 3·1. It then slowly decreases in brightness for 192 days to a minimum of 4·0 mag., at which it remains for 330 days. Then its brightness returns to normal in 192 days, and remains there for another quarter of a century. Except for the unusual period, there is nothing abnormal in this behaviour: it points simply to a double star whose components happen to be moving in a plane passing through or nearly through the earth, the flat minimum meaning that the eclipse is total, and the value of the magnitude change during eclipse suggesting that the components are not far different from one another in luminosity. Unfortunately, in this last respect the spectroscopic observations tell quite another tale. Spectroscopic binaries often show doubled spectra, each component having its own spectrum which indicates alternately recession and approach, by the Doppler shift of the lines. Many, however, show only one spectrum whose lines oscillate about their mean position. In this class falls ϵ Aur. But such a spectrum as this points unmistakably to a big difference of luminosity between the components, in point-blank disagreement with the photometric data. And this contradiction between the spectroscopic and the photometric results is not to be explained away as experimental error. The discrepancy is the more puzzling in that, taken separately, the two arguments are irreproachable: they have been used individually in hundreds of cases and are at the basis of much of our statistical knowledge of stellar duplicity; and they have been applied in conjunction many times before without leading to contradictory results.

A complete interpretation of the paradox has been given quite recently by a team of workers at Yerkes[14]. The model they postulate is of a character sufficiently unusual for accounts, more or less sensational, to have reached the daily press. It consists of a small, massive, bright *F*2 supergiant eclipsed every 27 years by a very much larger translucent component at a temperature of only about 1300° K. and of diameter 3000 times that of the sun. This large *I* ("infra-red") star will give no detectable spectrum in the region covered by photographic observation, and there will be in effect only one eclipse per revolution, as observed. The photometric results can now be explained. Throughout eclipse the light observed is practically that of the ***eclipsed F*** star, the diminution in brightness being caused by absorption in the atmosphere of the *I* star, and not, as had previously been tacitly assumed, by simple removal of the eclipsed star's light.

The argument leading to these results runs as follows. From simple geometrical consideration of the light curve, we can calculate, for various assumptions as to the circumstances of the eclipse, central, grazing or intermediate, the radii of the components in terms of the orbital dimensions. The orbital dimensions are determined from the radial velocity curve, which indicates an orbital velocity of 15 km./sec., and the period. Thence the radii are obtained in km. for these various assumptions. From the dimensions of the *F* star and a knowledge of its temperature from its spectral class, various values of its luminosity, corresponding to the various

initial assumptions, can be calculated. For the mass given by the dynamics of the orbit, the mass-luminosity relation gives a luminosity 6×10^4 times greater than that of the sun (a supergiant indeed!), and fixes the eclipse as intermediate between central and grazing occultation. The *I* star on this basis has a diameter 15 times that of the *F* component, and its luminosity fixes its temperature between 1200° and 1400° K.

The existence of such infra-red stars has often been suspected, but we know little or nothing of their properties. We do know, however, that the opacity of ordinary stellar material is a matter of ionization, controlled at any rate partially by the temperature. At a temperature of 1300° K., temperature ionization would be so low that the star might reasonably be transparent. If this idea is correct, the drop in brightness of the system when the *F* star goes into eclipse is likely to be a result of ionization opacity produced in the *I* atmosphere by ultra-violet radiation from the *F* component. (The enormous size of the *I* star makes it probable that its outer layers are almost as close to the *F* star as they are to its own centre.) The crucial test of this theory comes when it is asked to explain the flat minimum of the light-curve during totality. If the photoionization is produced uniformly throughout the *I* atmosphere, a curved minimum would result from the passage of light through a series of chords of increasing and then decreasing lengths. But, as the authors point out, the ultra-violet *F* radiation will be quickly absorbed in the outer layers of the *I* star. It will in effect produce a Heaviside layer of fairly constant thickness over the whole hemisphere presented to the *F* star. The opacity to visible light will be localized in this layer, the rest of the material in the infra-red component not contributing to what will therefore be almost constant absorption during the total phase.

Confirmation of the reality of this model comes from spectroscopic observations. During the eclipse the ionized atmosphere of the *I* star contributes its own lines, e.g. those of Fe II, Ti II and Ca II, to the composite spectrum, identical with those of the primary, but displaced by rotational Doppler effect. The *I* component is so large that its modest rotation produces an equatorial velocity of 50 km./sec., and the lines therefore have a red satellite during the early part of the eclipse and a violet one towards the end of the eclipse. In the portion of the *I* star presented to the earth, photoionization does not occur, and the resonance lines of neutral elements might be expected; the *D* lines of Na are indeed observed at high intensity during the eclipse.

Direct evidence for the existence of the *I* component has recently been obtained by Hall[15] from a spectrophotometric study of the star, using a photoelectric photometer. He finds that at long wave-lengths (8000 to 10,000 A.) the composite colour of the binary system is redder than that of a normal *F*2 supergiant, the excess energy reaching 0·13 mag. at 9600 A. Kuiper[16], using this value for the excess, has calculated the physical properties of the two components with results in good agreement with the model postulated. Finally, three Harvard astronomers[17] have measured the radiometric magnitude of the system as 2·65. The visual magnitude of the binary, and therefore of the *F* star too, since the *I* star is ineffective visually,

is 3·16. The average heat index of *F*2 stars being 0·25, this yields a radiometric magnitude of 2·91 for the *F* component. Using the Yerkes effective temperatures and radii-ratio, the radiometric magnitude of the *I* star is calculated to be 4·20. The total radiometric magnitude expected for the system, then, is 2·62, in excellent agreement with observation. There seems little doubt that the ϵ Aur system is a mystery no longer.

REFERENCES

(1) Rosenthal, A. H. *Z. Astrophys.* **1**, 115 (1930).
(2) Goudsmit, S. and Wu, Ta-You. *Astrophys. J.* **80**, 154 (1934).
(3) Waldmeier, M. *Z. Astrophys.* **15**, 44 (1938).
(4) Pettit, E. *Publ. Astr. Soc. Pacific*, **50**, 168 (1938).
(5) Moore, J. H. *Publ. Astr. Soc. Pacific*, **35**, 333 (1923).
(6) Greenstein, J. L. *Astrophys. J.* **87**, 151 (1938).
(7) Greenstein, J. L. *Circ. Astr. Obs. Harv.* **422** (1938).
(8) Williams, E. G. *Astrophys. J.* **79**, 280 (1934).
(9) Heard, J. F. *J. R. Astr. Soc. Canada*, **31**, 60 (1937).
(10) Greaves, W. M. H. and Martin, E. *Mon. Not. R. Astr. Soc.* **98**, 434 (1938).
(11) Greaves, W. M. H. and Martin, E. *Observatory*, **60**, 67 (1937).
(12) Woolley, R. v. d. R. *Mon. Not. R. Astr. Soc.* **95**, 115 (1934).
(13) Gussow, M. *Veröff. Berlin-Babelsberg*, **11**, H. 3 (1936).
(14) Kuiper, G. P., Struve, O. and Strömgren, B. *Astrophys. J.* **86**, 570 (1937).
(15) Hall, J. S. *Astrophys. J.* **87**, 209 (1938).
(16) Kuiper, G. P. *Astrophys. J.* **87**, 213 (1938).
(17) Emberson, R. M., Loevinger, R. and Sterne, T. E. *Bull. Astr. Obs. Harv.* **908**, 11 (1938).

V. THE SPECTRA OF THE ALKALI-METAL HYDRIDES

By R. W. B. PEARSE, Ph.D., F.R.A.S.
Imperial College, London

§1. INTRODUCTION

During the year accounts have been published of investigations of the spectra of the diatomic alkali-metal hydrides, RbH and CsH. Since data are now available for all of the five members of this series of molecules, LiH[1], NaH[2,3], KH[4], RbH[5], CsH[6], a brief survey of the spectra of the group will perhaps not be out of place in this year's *Reports*. These spectra are of interest from several points of view. In the first place they exhibit features not found so far in the spectra of any other diatomic molecules. Then, their open structure, combined with the fact that an unusually large number of bands are readily observed in each system, has enabled the energy terms to be examined in more than ordinary detail. Moreover, comparison with the corresponding deuterides enables a rigorous test to be made of the mechanical significance attached to the constants of the energy terms. The spectra provide

a good example of the similarity of the energy levels and band systems of molecules formed from atoms closely related in the periodic table. Finally, it may be noted that they are of interest from the theoretical point of view in that they belong to a group of molecules that are among those having the simplest electronic configurations, molecules formed from two atoms each possessing a single valence electron.

§2. GENERAL DESCRIPTION

In appearance and in method of production the spectra of the alkali-metal hydrides are remarkably alike. They may be observed in absorption by heating the appropriate metal in an atmosphere of hydrogen at a suitable pressure and in emission from the arc in hydrogen or from a discharge tube containing hydrogen and the vapour of the metal. Over a range of some 2000 A. in or near the visible region each shows a great number of lines without heads or regularities obvious to casual inspection. The spectra in this are very similar to the spectrum of the hydrogen molecule, which indeed, from many points of view could be included in this group. As the atomic weight of the alkali atom increases, the region occupied by these lines gradually shifts from the violet towards the red. The ranges recorded by various observers are: LiH: 3000–4500 A., NaH: 3500–5000 A., KH: 4100–6600 A., RbH: 4800–6500 A., CsH: 4500–6250 A.

Analysis shows that in each case the spectrum consists of from twenty to thirty bands degraded to the red, each band possessing a single P and a single R branch. The R branches turn back so sharply that the heads, being formed from weak lines near the origins of the bands, are not conspicuous. The complexity of the spectra arises from the great amount of overlapping of these simple bands which takes place. In general, about half the bands of each spectrum are found to fall into a long $(v', 0)$ progression while of the remainder the majority belong to the $(v', 1)$ and $(v', 2)$ progressions. The electronic transition is found to be of the type $^1\Sigma \rightarrow {}^1\Sigma$, the peculiar structure of the bands and their unusual distribution arising from the smallness of the molecular constants in the upper states as compared with those of the lower states. Closer examination of the energy terms of the upper states shows that these are also of unusual form.

§3. THE EMPIRICAL ENERGY TERMS

As usual, the total energy of the molecule in wave-number units, E/hc, is considered as the sum of three terms:

$$E/hc = T_e + G(v) + F_v(J),$$

where $$G(v) = \omega_e (v+\tfrac{1}{2}) + x_e\omega_e (v+\tfrac{1}{2})^2 + y_e\omega_e (v+\tfrac{1}{2})^3 + z_e\omega_e (v+\tfrac{1}{2})^4 + \ldots$$

and $$F_v(J) = B_v J(J+1) + D_v J^2(J+1)^2 + H_v J^3(J+1)^3 + \ldots.$$

If for the energy of vibration $G(v)$ we write $\omega_v (v+\tfrac{1}{2})$, where ω_v is interpreted as the frequency of vibration for state v, then

$$\omega_v = \omega_e + x_e\omega_e (v+\tfrac{1}{2}) + y_e\omega_e (v+\tfrac{1}{2})^2 + z_e\omega_e (v+\tfrac{1}{2})^3 + \ldots.$$

In the case of most molecules the coefficients $x_e, y_e, z_e \ldots$ decrease in this order and

x_e is negative so that the frequency of vibration decreases with increasing energy from its value ω_e for infinitesimally small vibrations to the value zero at dissociation. The coefficients in the rotational energy term $F_v(J)$ depend on the state of vibration considered; this dependence can be described by the supplementary formulae

$$B_v = B_e + \alpha_1 (v+\tfrac{1}{2}) + \alpha_2 (v+\tfrac{1}{2})^2 + \ldots,$$
$$D_v = D_e + \beta_1 (v+\tfrac{1}{2}) + \beta_2 (v+\tfrac{1}{2})^2 + \ldots,$$
$$H_v = H_e + \gamma_1 (v+\tfrac{1}{2}) + \gamma_2 (v+\tfrac{1}{2})^2 + \ldots,$$

where $\alpha_1 \gg \alpha_2$, etc. In the above expressions a slight but important departure has been made from the convention hitherto adopted, namely the use of positive instead of negative signs for the coefficients $x_e\omega_e$ and α_1, so that positive signs now appear throughout. This is especially convenient for the present discussion and could be adopted generally with advantage. Except for some minor corrections to be mentioned later $B_v = h/8\pi^2 c\mu r^2$, μ being the reduced mass of the molecule and r the distance between the nuclei. For most molecules α_1 and D_v are negative. This implies that the asymmetry of the vibrations and the centrifugal effect of the rotation causes the average value of r to increase. The energy terms of the upper electronic state of each of the alkali-metal hydrides are exceptional in that x_e and α_1 are positive, so that ω_v and B_v at first increase with v, attain maximum values and then decrease regularly. Thus in their excited states these molecules appear to shrink in size as well as to vibrate more rapidly as the amount of vibrational energy increases.

The available data are collected in table 1. The following remarks apply to all of the molecules. Owing to the grouping of the bands into three fairly extensive progressions, namely the $(v', 0)$, $(v', 1)$ and $(v', 2)$, values of the constants for the three lowest vibrational levels of the ground state may be determined from a large number of bands and so are correspondingly accurate; but since higher levels are not observed, the coefficients of the higher powers of $(v''+\frac{1}{2})$ cannot be determined. In the excited state on the other hand, a large range of values of v' is available, but since the bands for the lowest values are extremely weak there is usually uncertainty in the assignment of absolute values to their respective bands. This means that the energy terms may be determined to high powers of $(v'+\frac{1}{2})$ but are subject to the possible necessity for a change of origin. This doubt may be removed by comparison with the corresponding deuteride, but this has only been done for lithium and sodium.

The data are most extensive and accurate for LiH as a result of the careful and detailed analysis of this spectrum and the corresponding deuteride by Crawford and Jorgensen[1]. The values for NaH are not very satisfactory. The spectrum was first studied in absorption by Hori[2]; later Hori made a more extensive study of the emission spectrum, as a result of which he increased the values of v' by one unit; still more recently Olsson[3] has observed the spectrum of the deuteride in absorption and shown the necessity for a further increase of three units in v'. Data are thus

wanting for the lowest three vibrational levels of the upper state. The results shown in the table have been obtained by extrapolation from Hori's data. This is none too safe as the constants for LiH change very rapidly at the three lowest values of v, and it is probable that those for NaH will behave in a similar way though to a lesser degree. The data for KH obtained by Almy and Hause[4] are more complete, and since the distribution of bands observed is very similar to that found for LiH and corrected NaH the v' assignment is probably correct. The constants given for RbH are

Table 1. *Observed constants for the band spectra of the alkali-metal hydrides*

Constant	LiH	NaH	KH	RbH	CsH
Ground state $x^1\Sigma^+$					
Electronic term T_e''	0	0	0	0	0
Vibrational term $G''(v)$					
ω_e	1405·7	1172·9	987·4	936	890·7
$\omega_e x_e$	−23·20	−20·3	−15·5	−14	−12·6
$\omega_e y_e$	0·16	$0·1_8$	$0·0_9$	—	$0·1_1$
Rotational term $F''(J)$					
B_e	7·5131	4·9012	3·415	3·019	2·709
α_1	−0·2132	−0·1353	−0·083	−0·07	−0·057
$\alpha_2 \times 10^4$	7·5	—	—	—	—
$D_e \times 10^4$	−8·617	−3·32	−1·65	−1·2	−1·00
$\beta_1 \times 10^4$	−0·159	—	—	—	—
$H_e \times 10^8$	11·4	1·3	1·5	—	—
Excited state $A^1\Sigma^+$					
Electronic term T_e'	26,516	22,724	19,525	18,912	18,410
Vibrational term $G'(v)$					
ω_e	234·41	308·6	254·5	248	204·0
$\omega_e x_e$	28·95	5·5	3·26	5	5·70
$\omega_e y_e$	−4·18	−0·2	+0·007	−0·3	−0·35
$\omega_e z_e$	+0·51	—	−0·01	—	+0·01
Rotational term $F'(J)$					
B_e	2·8186	1·756	1·344	−1·234	1·126
α_1	0·0783	0·057	0·030	−0·019	0·0185
α_2	−0·0263	−0·005	—	−0·002	−0·00246
$D_e \times 10^4$	−16·3	−2·15	−1·44	—	−1·04
$H_e \times 10^9$	100	—	3·8	—	—

from work in progress by Gaydon and Pearse[5]; the spectrum has been observed very extensively in emission and the distribution is very similar to that for KH. The values for CsH are obtained from the recent paper by Almy and Rassweiler[6]; in this case the spectrum was observed in absorption with the consequent disadvantage that only the progressions $(v', 0)$ and $(v', 1)$ were obtained so that the energy terms for the ground state are incomplete.

A glance at the results in table 1 shows that for the ground state the constants decrease smoothly from LiH to CsH. The same trend is observable in the electronic energy of the excited state. The first vibrational constant ω_e shows a very different behaviour, rising to a maximum for NaH. The irregularities in the smaller constants are not to be taken too seriously without further investigation, as these constants are very sensitive to details of computation such as the number of terms used.

§4. THE DERIVED MOLECULAR CONSTANTS

In theoretical treatments of the rotating-vibrating diatomic molecule it is usual to assume that there is some stable equilibrium value of the internuclear distance, r_e, and to express the dependence of the potential energy of the molecule on the internuclear distance at values near this by a plausible function. Many such functions have been proposed but for small displacements from the equilibrium position only one will be considered here, namely

$$U(r) = a_0\xi^2(1 + a_1\xi + a_2\xi^2 + \ldots),$$

where $\xi = (r - r_e)/r_e$. Using this potential energy function the various coefficients may be evaluated in terms of the observed band spectrum constants. The necessary relations are

$$a_0 = \omega_e^2/4B_e,$$
$$a_1 = -(1 - \alpha_e\omega_e/6B_e^2),$$
$$a_2 = 5a_1^2/4 + 2x_e\omega_e/3B_e.$$

The equilibrium value of the internuclear distance may be obtained from the relation

$$B_e = h/8\pi^2 c\mu r_e^2,$$

since the value of the reduced mass is known from the atomic weights and the mass of unit atomic weight. The moment of inertia I_e is simply μr_e^2. The values of the derived constants listed in table 2 are obtained from these relations as they stand. It should be noted, however, that the usual treatment of the diatomic molecule neglects many effects which change the values of the constants of the energy terms by small, but nevertheless measurable, amounts. The failure of the simple isotope theory to describe the effect completely for hydride molecules is evidence for this. Crawford and Jorgensen have listed four of these effects and give references to the literature on each. The effects are:

(1) Anharmonicity terms due to the interaction of rotation and vibration. The correction terms arising from this are of importance only for hydrides and other light molecules.

(2) Terms due to the interaction of electronic and nuclear motions (neglecting the interaction of neighbouring electronic levels).

(3) L uncoupling terms. These are due to the interaction of electronic states whose Λ values differ by ± 1.

(4) Terms arising from the interaction of states of the same Λ value.

Using the results of Dunham's calculations the above mentioned authors investigate the first effect in detail for LiH and LiD; for LiH the results are:

	B_e'	ω_e'	B_e''	ω_e''
Uncorrected	2·8186	234·41	7·5131	1405·649
Corrected	2·8375	236·225	7·5149	1405·401

Evaluation of the other effects cannot be carried through in the absence of more detailed knowledge of the electronic states of the molecule.

In cases where long progressions of bands are obtained a very good estimate of the energy of dissociation of a molecule may be obtained by extrapolating the expression for the vibrational energy to its first maximum value. If the following coefficients are small a first approximation value may be obtained from the leading two, ω_e and $x_e\omega_e$. The energy of dissociation $\boldsymbol{D}_e$ is equal to $\omega_e^2/4x_e\omega_e$. This can be applied to the ground states of the alkali-metal hydrides although the extrapolation is somewhat precarious in view of the short range actually known. This method cannot be applied to the excited state, however, since the value of $x_e\omega_e$ has a negative sign. In view of the anomalous nature of the vibrational energy term the energy of dissociation of the upper state has usually been evaluated in another way. If it is assumed that the molecule dissociates into a normal hydrogen atom and an alkali atom in its first excited ^{2}P state, then

$$\boldsymbol{D}_e' = (^2\mathrm{P} \rightarrow {}^2\mathrm{S}) - (T_e' - T_e'') + \boldsymbol{D}_e''.$$

In table 2 are collected the values of the derived constants calculated from the data in table 1. It will be noticed that the values of the force constants, a_0, a_1, a_2, for the ground state are very closely alike for all the molecules but especially so for

Table 2. *Derived molecular constants*

Constant	LiH	NaH	KH	RbH	CsH
Ground state $X^1\Sigma^+$					
$r_e \times 10^8$ cm.	1·59	1·88	2·24	2·36	2·49
$I_e \times 10^{40}$ g. cm^2	3·68	5·64	8·10	9·16	10·21
$a_0 \times 10^{-4}$ cm.$^{-1}$	6·58	7·02	7·14	7·25	7·32
a_1	−1·88	−2·10	−2·17	−2·2	−2·15
a_2	2·38	2·76	2·87	2·7	2·69
$\boldsymbol{D}_e$ e.v.	2·63	2·09	1·94	1·9	1·94
Excited state $A^1\Sigma^+$					
$r_e \times 10^8$ cm.	2·59	3·15	3·56	3·69	3·86
$I_e \times 10^{40}$ g. cm.	9·81	15·75	20·58	22·41	24·56
$a_0 \times 10^4$ cm.$^{-1}$	0·487	1·36	1·20	1·24	0·924
a_1	−0·615	−0·049	−0·295	−0·48	−0·504
a_2	7·32	2·09	1·73	3·0	3·69
$^2P_{\frac{3}{2}} \rightarrow {}^2S$	14904	16956	12985	12579	11178
$^2P_{\frac{1}{2}} \rightarrow {}^2S$	14904	16973	13043	12816	11732
$\boldsymbol{D}_e$ e.v.	1·20	1·38	1·13	1·1	1·05
Atomic ionization potential	5·37	5·12	4·32	4·16	3·87

the last four in the series. In this connexion it may be remarked that the constancy of a_0 proved valuable in the analysis of RbH, as it enabled a very accurate estimate of B_e to be made from the vibrational analysis to serve as a guide in the rotational analysis. In the excited state the values are much more variable. The first force constant a_0 reaches a maximum value at NaH in accordance with the maximum of ω_e. The behaviour of the energies of dissociation is interesting. In the ground state LiH appears to be the most stable and CsH probably the least stable, although the drop after NaH is very small. In the excited state NaH appears to be the most

stable and CsH again the least. The rise to this maximum value at NaH may be correlated with the behaviour of the atomic energy levels of the alkali atoms themselves. It will be noticed from table 2 that the interval $^2P \rightarrow {}^2S$ is a maximum at Na, so that if the excited state of the NaH molecule is to dissociate giving a ^2P Na atom the quantity $(T_e' - T_e'') + \boldsymbol{D}_e' - \boldsymbol{D}_e''$ must be a maximum for NaH. It seems unlikely that the energy of dissociation of NaH in the ground state should be exceptionally small, so the extra energy must be shared in some way between the energy of dissociation and the electronic energy of the excited state. It is evident from table 2 that it is the energy of dissociation which is large, and with this may be coupled the large value of ω_e in table 1.

The unusual behaviour of the constants ω_v and B_v may now be examined in terms of the potential energy function. If this is plotted for the region where $r = r_e$, then normally the curve is asymmetrical; the side for r less than r_e rises more rapidly than a parabola while the side for r greater than r_e falls below. This asymmetry is usually such that as the vibrational energy increases ω_v decreases and the average value of μr^2 increases making α_1 negative. In the upper states of the alkali-metal hydrides the $U(r)$ curve is found to rise at first more rapidly than a parabola on both sides of r_e thus causing the initial increase in ω_v. The average value of μr^2 decreases making α_1 positive. It remains to be explained why the alkali-metal hydrides should have $U(r)$ curves differing from all other molecules in this particular way. This requires some discussion of the nature of the electronic energy states of these molecules.

§5. THE ELECTRONIC ENERGY STATES

The two electronic states observed are both $^1\Sigma$. The possible states which may be derived from a hydrogen atom in its normal state and a sodium atom are $^1\Sigma$, $^3\Sigma$ for the ^2S state of sodium and $^1\Sigma$, $^3\Sigma$, $^1\Pi$, $^3\Pi$ for the ^2P state. The observed states have been identified with the two predicted $^1\Sigma$ states. There is, however, in this view of the nature of these states nothing peculiar which would suggest that their potential energy functions should be unusual in form. In a recent series of three papers Mulliken[7] has discussed in detail the lowest energy states of molecules formed from atoms which have either one electron in excess of a closed shell or one vacancy. The behaviour of these states with regard to such characteristics as polarity and ionicness is compared for the hydrogen molecule, the alkali-metal molecules, the alkali-metal hydrides and the hydrides and halides of the metals of the silver group. In this comparison it is noted that for the alkali-metal hydrides the energy of the combination $(M^+ + H^-)$ is not as large as usual compared with $(M + H)$, where M is taken to represent one of the alkali atoms. The lowest state of the molecule formed from the two ions is also $^1\Sigma$. Mulliken concludes that the exceptional shape of the upper potential energy curve can be explained if it is supposed that the true curves of the molecule are the result of a strong interaction between the ionic curve and the previously considered atomic curves. The case of LiH is treated in most detail. It is shown that the upper state appears to be only partially

ionic near its minimum but at slightly larger r values goes over into the pure ionic curve. The lower state, on the other hand, is strongly polar near its minimum but goes over into a pair of neutral atoms with increasing r. Rosenbaum [8] has examined this suggestion a little more closely by drawing the potential curve for the upper state of LiH from observed band spectrum data and comparing it with the Coulombic curve for $(Li^+ + H^-)$. He concludes that the curve does not go over asymptotically into the ionic curve but actually crosses it at somewhat smaller internuclear distances finally going over to Li (^2P)+H. This conclusion does not invalidate Mulliken's interpretation of the nature of the state, as in either case it seems that the state must be predominantly ionic in character over a range of moderately large r values before dissociating into neutral atoms. The other alkali-metal hydrides have not been considered in detail but similar relations undoubtedly hold.

REFERENCES

(1) Crawford, F. H. and Jorgensen, T. *Phys. Rev.* **47**, 358 (1935); **47**, 932 (1935); **49**, 745 (1936).
(2) Hori, T. *Z. Phys.* **62**, 352 (1930); **71**, 478 (1931).
(3) Olsson, E. *Z. Phys.* **93**, 206 (1935).
(4) Almy, G. M. and Hause, C. D. *Phys. Rev.* **42**, 242 (1932).
(5) Gaydon, A. G. and Pearse, R. W. B. *Nature, Lond.*, **142**, 291 (1938).
(6) Almy, G. M. and Rasswеiler, M. *Phys. Rev.* **53**, 890 (1938).
(7) Mulliken, R. S. *Phys. Rev.* **50**, 1017 (1936); **50**, 1028 (1936); **51**, 310 (1937).
(8) Rosenbaum, E. J. *J. Chem. Phys.* **6**, 16 (1938).

THE SOFT X-RAY SPECTROSCOPY OF THE SOLID STATE

By H. W. B. SKINNER
H. H. Wills Physical Laboratory, University of Bristol

§ 1. INTRODUCTION

The study of X-ray spectra has led to results of first class importance in two branches of physics, namely, in the elucidation of the lattice structure of crystals, and of the inner-shell structure of atoms. But only recently have attempts been made to use it in order to build up a spectroscopy of the solid state, that is to say to determine the electron level structure characteristic of a given solid in much the same way that the level structure of an atom is determined in optical spectroscopy.

Since the deeper-lying levels are not appreciably affected by the lattice-binding, it is the valence-electron leveis in which we are interested, so one might ask at the outset: "Since atomic spectroscopy has been based on spectra in the visible and near ultra-violet regions, why should X rays be required for the solid spectroscopy?" The fact is that the long wave-length emission and absorption spectra of a solid are of little use in determining the level system of its valence electrons. Such spectra in the visible and ultra-violet regions of the spectrum exist. For example, the true absorption of a thin metal foil can be disentangled from the mixture of absorption and selective reflection which causes its colour by transmitted light.* On the other hand, the emission spectra of metals, though little known in comparison with the absorption spectra, have been observed in a number of cases by Mohler and Boeckner[2]. There are technical difficulties here, because, as is well known in experiments on the fluorescence of crystals, etc., the results can be entirely dependent on small traces of impurity, and even the presence of a gas layer on a baked-out metal surface in a good vacuum can serve as such an impurity. The experiments of Mohler and Boeckner, performed with great care, are probably free from such criticism. But from the point of view of solid spectroscopy, the results are disappointing; both the emission and absorption spectra of pure solids† appear as broad bands stretching through large regions of the spectrum without any sharp characteristics.

The reason for this absence of characteristics in the optical emission and absorption spectra of a solid lies in the nature of the level system of its valence electrons. The optical spectra may be defined as corresponding to transitions of an

* For a recent account, see O'Bryan[1].

† One excepts such cases as those of the rare-earth salts where the absorption is due to the presence of an incompleted inner shell. Also the cases of the alkali halides which Hilsch and Pohl[3] and Schneider and O'Bryan[4] have proved to show a type of line absorption. The fact is that the halides seem to possess much sharper valence levels than other solids (see below, § 3 (*f*)).

electron from one valence level to another. Now it is well known that the level system of a valence electron in a solid forms a continuum; that is the energy levels are not discrete like those of a free atom. This is illustrated in figure 1, in which the spectrum emitted by solid aluminium in the region of 100 to 200 A. is shown side by side with the spectrum of a vacuum spark between Al electrodes, which gives lines emitted by the vapour.

Corresponding to every single level of the free atom, we have in the solid a number of levels equal to the total number of atoms contained in the piece under consideration; this practically infinite number of levels is spread out by interaction between the atoms into a band of levels of finite energy breadth. Since the bands corresponding to different atomic levels overlap, it follows that the complete level system of the lattice extends continuously into the region of high energies and can only be specified by a density function $N(E)$, where $N(E)\,dE$ is the number of levels (per atom) between E and $E+dE$. It is this function $N(E)$, defining the

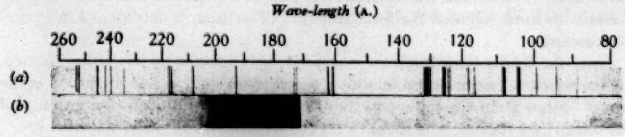


Figure 1. Emission spectra of Al. (*a*) from vapour; (*b*) from solid.

level system of the solid, which it is our problem to determine spectroscopically. In general, it is a complicated curve, which, as we shall show by experiment, is by no means devoid of characteristic features which may even be quite sharp.

Our present purpose is to show that it cannot be determined from the optical spectra of the solid. This is due to the fact that the energies of the valence electrons of a solid in the lowest state are not all the same. For, according to the Pauli principle, a level of a given type can be occupied only by two electrons (with opposite spins). The available valence electrons therefore fill the level system up to a certain maximum energy. For metals, the energy spread of this band of *occupied* levels lies between about 3 and 15 ev. in different instances, and most insulators seem to have an even wider band.

We thus see that the optical emission and absorption spectra of a solid correspond to transitions between two sets of levels, each of which is a part of the $N(E)$ curve and neither of which is sharply defined in energy. Although there arises an extra selection rule (the k selection rule), yet it remains true that there is no one-to-one correspondence between an emitted or absorbed wave-length and a definite pair of levels in the lattice-level system. It follows that the characteristics of the valence-electron level system of a solid cannot be disentangled from the optical emission and absorption spectra.

But if we pass to the case of X-ray spectra, it is easy to see immediately that the problem of determining the $N(E)$ curve for a solid can be solved, at least in

principle. Unlike the valence electrons in a crystal, the electrons in the *inner shells* of the atoms are more or less bound to their individual atoms, instead of running throughout the crystal. The interaction between the lattice and these inner shells is weak, and therefore the corresponding energy levels remain fairly sharp. In principle they are spread out into a band, but the absolute width of this band is very small. We may therefore regard these X-ray levels as being only slightly broadened by the lattice field.

It will at once be apparent that these sharp X-ray levels are exactly what we require for our problem of the determination of $N(E)$. The problem divides itself

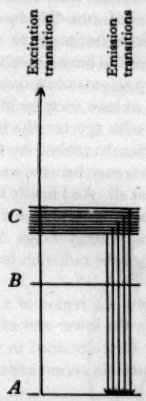


Figure 2. X-ray emission process.

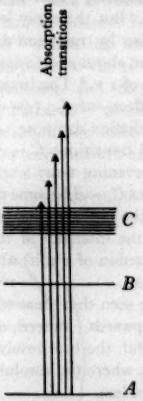


Figure 3. X-ray absorption process.

naturally into two experimental problems, those of emission and absorption. In the first, figure 2, an electron is ejected by some method (such as by electron impact in an X-ray tube) from a level of an inner shell (level A). Into the vacancy, transitions of an electron from the continuous band of filled valence levels may occur.* These will give a continuous band of radiation which will reflect accurately the characteristics of the $N(E)$ curve over the region of *filled* levels. For, before the emission process, the lattice has all its normally occupied states filled, except that there is a vacancy in the level A. The energy of the initial state is therefore sharply defined within the limits of the breadth of the level A. After the emission, an electron may be missing from any of the levels of the filled valence-electron band. The corresponding absorption problem is even simpler: we determine the probability of a transition from the level A, figure 3, into one of the *vacant* levels.

* Transitions of the type $B \rightarrow A$ are frequent in soft X-ray spectra. They may be classed as internal atomic transitions, and since they depend only slightly on the lattice binding, they are of little interest from the point of view of this article.

The characteristics of the $N(E)$ curve for the *vacant* levels of the lattice will obviously be reflected in this absorption curve. In both emission and absorption experiments a factor involving the absolute transition probability enters, but we shall leave over the discussion of this question until § 3 (*d*).

We have shown that X-ray spectra can give information of the type required for a solid, while the optical spectra fail. It remains to consider what are the most suitable X-ray spectra for the purpose. Various considerations enter here. The most obvious is that the X-ray level must be sufficiently sharp for the purpose; in fact the reciprocal of the energy breadth of the X-ray level (ΔE) may be defined as the theoretical *resolving power* of the method as applied to the determination of the $N(E)$ curve. But the X-ray level is effectively broadened, not only by the lattice field, but also by radiation damping. This factor alone limits us to the region of X rays from about a few angstroms upwards, if we wish to have a resolving power of the order of 1 V.* The broadening is augmented by an extra broadening due to the Auger effect—see § 3 (*e*)—which, in certain cases, at least, may be greater than that from radiation damping. Also, in the case of emission spectra, the intensity of the required transitions $C \rightarrow A$, figure 2, may be seriously robbed by transitions from an intervening X-ray level B; in extreme cases this may become so acute that the transitions $C \rightarrow A$ become very difficult to observe at all. And finally there is the question of the experimental resolving power; it is clearly more difficult in general to measure the intensity of the radiation over a small energy range ΔE (of the order of a fraction of a volt) when the absolute energy of the radiation is large (e.g. 4000 V. at 3 A.).

It will be seen that these considerations drive us into the region of X rays from about 5 A. upwards. Indeed, although results towards the lower end of the range are very useful, the best resolving power has actually been obtained in the region above 100 A., where the absolute breadth has been reduced in recent experiments to 0·1 V. or even less.

The difficulties in the use of these soft X-ray spectra are entirely technical and have been overcome in recent years, very largely through the work of Prof. Siegbahn.

§ 2. TECHNICAL METHODS

(*a*) *Spectroscopic technique.* The region of the spectrum in which we are most interested may thus be described as stretching from about 5 to 500 A. This region covers the K radiation of elements Li (3) to S (16), the L radiation of elements Na (11) to Zn (40), and the M radiation of elements K (19) to Hf (72). The whole

* For hydrogen-like atoms with a variable nuclear charge, a simple calculation gives the breadth associated with the $2p \rightarrow 1s$ transition as proportional to ν^3, where ν is the frequency of the transition. Assuming the value for hydrogen itself, we obtain the following values of the breadth ΔE corresponding to various wave-lengths:

	1000	100	10	1	A.
ΔE	4×10^{-6}	4×10^{-4}	0·04	4	V.

Of course, the results will vary considerably with the type of transition involved[44].

section of the spectrum consists of radiation very absorbable in air, and vacuum spectroscopy is essential.

The region falls naturally into two parts: a region of crystal spectroscopy and a region of grating spectroscopy. The limit between the two is sharply set by the extreme wave-length diffracted by mica (19·9 A.), since it has not so far been found possible to obtain a satisfactory crystal with a larger spacing. Where the crystal can be used, it is superior to the grating both in resolution and in intensity.

In the crystal region, one may employ the normal methods of X-ray spectroscopy using a slit and a crystal rocked through definite angle limits. This has in fact been used very successfully, for example, by Saur[5] and Gwinner[6] in work to which we shall subsequently refer on the *L* radiation of elements Fe to Ge. The alternative is the curved crystal method of Johann[7] which has been used, for example, by Sandström[8] in experiments on *L* absorption of the elements Cu to Pd and the *K* absorption of Mg and Al; by Karlsson and Siegbahn[9] and Farineau[10] on the *K* emission of Mg and Al; and by Farineau[10] on the *L* emission of elements Fe to Zn. The method gives considerably more intensity than the straightforward method, and since it is little known, a brief account may not be out of place here. It is illustrated in figure 4.

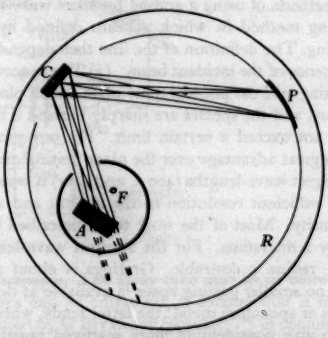


Figure 4. Bent crystal spectrograph. *C*, crystal; *A*, anticathode; *F*, X-ray filament; *P*, photographic film; *R*, Rowland circle.

The crystal is bent to a certain radius of curvature (say from 50 cm. upwards). The photographic plate or film is bent onto the Rowland circle, that is a circle of half that radius of curvature. The X rays fall, without passing a slit, on the crystal. Those which strike the crystal at an angle corresponding, according to the Bragg relation, to their wave-length, are reflected and *focused* at a point on the photographic film. Thus a given setting of the anticathode in position relative to the crystal gives, assuming that the corresponding wave-lengths are being emitted from it, a certain wave-length stretch of the spectrum. A movement of the anticathode

shifts the observed wave-length region. It will be seen that the X-ray tube has to be built into the spectrograph container. This perhaps is of some disadvantage in comparison with the ordinary method using slits, on the grounds of inaccessibility.

The methods (and results) of grating spectroscopy in the soft X-ray region have been described by Siegbahn in a recent article[11], and we shall give only a brief account here.

The essential for the use of a grating for short wave-length spectroscopy is that the X rays must be allowed to fall on it at a small glancing angle; this angle must be about equal to the angle of reflection of radiation entering the material of the grating (the refractive index of which for these short wave-lengths is, of course, less than unity). The glancing angle may, for example, vary between about 1° and 10° or more according to the region of the spectrum to be investigated. For the very shortest wave-lengths (20 A.) can be diffracted with a 5° angle of incidence, but the intensity is considerably greater if the angle is about 1°. A 5° angle may be used conveniently from about 50 A. upwards; but the use of a larger angle for wave-lengths above about 200 A. would have the advantage that the high order spectra of short wave-length would be weaker, and so less likely to confuse the spectrum to be observed.

There are two methods of using a grating for short wave-length spectroscopy: (1) The plane-grating method in which a beam defined by slits is allowed to impinge on the grating. The definition of the line then depends on the narrowness of the angle of divergence of the incident beam. (2) The concave grating method in which a slit, the grating and the photographic plate are all placed on the Rowland circle in the usual way, and the spectra are sharply focused if the angle of aperture of the grating does not exceed a certain limit. For our problems, the concave grating method has a great advantage over the plane grating method. Except in the region of the very longest wave-lengths (400 A. and over) it would be difficult, using the latter, to obtain sufficient resolution in the spectra and at the same time to retain sufficient intensity. Most of the work to be described has been performed with gratings of 1 or 2 m. radius. For the shortest wave-lengths (20 to 50 A.) a grating of 5 or 7 m. radius is desirable. Gratings of about 12,000 lines/cm. are generally used since no smaller grating space is available at the present time. The material may be glass or speculum metal: the latter tends, when its polish becomes slightly tarnished, to give considerably more scattered radiation; so glass is definitely to be preferred. The grating may often have its brightness considerably increased by slight etching with hydrofluoric acid. The optimum width of a grating is a function of the wave-length and angle of incidence and the radius of the grating. Roughly speaking, about 1 cm. of surface is required in most cases.

A schematic diagram of the concave-grating spectrograph used by Skinner and Johnston[12] for absorption work is shown in figure 5. The absorption technique will be described below (§ 2 (*d*)).

(*b*) *Photography*. Towards the short wave-length end of our spectrum region, up to about 50 A., ordinary photographic emulsions may be used for recording the radiation. But with increasing wave-length, they become slow, owing to the

absorption of the gelatine. Therefore one has to use special emulsions, and much work has been done with Schumann emulsions. But, recently, Messrs Ilford have developed a new type of emulsion, known as a "Q emulsion", which is intermediate in properties between the ordinary plate and the Schumann plate. This emulsion is very even in its properties and therefore is quite suitable for accurate photometry, including the determination of intensities. Except possibly in the region above 500 A., where its speed is considerably reduced on account of gelatine absorption, it is superior in properties to the Schumann plate. It is also faster than the ordinary plate, at least down to 20 A.

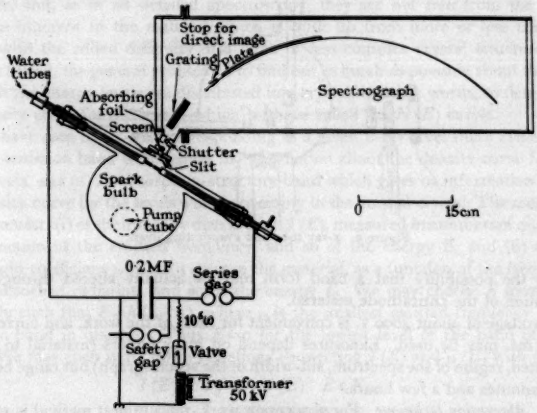


Figure 5.

(*c*) *X-ray tube technique.* The X-ray tube may be of more or less usual design, but since no windows can be employed, it follows that it must be continuously pumped. The bulb used by O'Bryan and Skinner[13]* in conjunction with a grating-spectrograph is illustrated in figure 6. The filament *C* with focusing shields stands opposite the anticathode *A*, a square sectioned copper tube drilled for water cooling which can be rotated on a cone joint, so as to be available for use on all four sides. Thus the four sides may be used in turn. The material to be investigated may be dusted on; but, actually, wherever possible was evaporated on from the evaporating unit *E*, in which a small piece may be heated by electron impact (as in figure 6) or otherwise. This has the advantage that the surface is laid down clean, and is easily replaced or interchanged. A glass X-ray tube was found more satisfactory than a metal X-ray tube in order to obtain the necessary vacuum (of the order of 10^{-5} to

* The technique of Siegbahn and his school is described in the article[11].

10^{-6} mm. of Hg) quickly and easily, and has the added advantage that one can watch the anticathode surface. A fast diffusion pump of speed about 10 litres/sec. was used for exhaustion in conjunction with a liquid air trap. In this way it was possible to obtain the required X-ray spectrum completely free from contamination by carbon or oxygen. Carbon can be specially troublesome, since the *K* radiation (44 A.) is diffracted in many orders by the grating, and it frequently happens that one of the orders falls on the band to be investigated. There is also

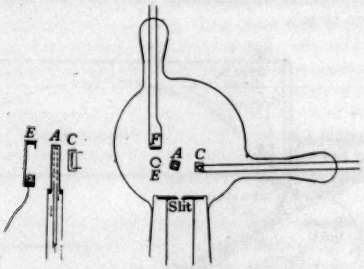


Figure 6. X-ray tube and evaporating oven.

always the possibility that a band form may be actually altered through contamination of the anticathode material.

A voltage of about 4000 V. is convenient for most of the work, and currents up to 200 ma. may be used. Exposures depend on many factors (material to be investigated, region of the spectrum, slit-width of the spectrograph) but range between a few minutes and a few hours.

(*d*) *Absorption technique.* For absorption work, the simplest method is to use a continuous spectrum which passes through a foil of suitable thickness of the material to be investigated. In the short wave-length region (near 10 A.) this method was used by Sandström[8] who made his foils by evaporating the metal, the absorption of which was required, onto a thin sheet of Al of negligible absorption in the region (somewhat to the long wave-length side of its *K*-absorption edge).

There are two difficulties in extending this method into the region of long wave-lengths (e.g. about 200 A.). The first is that the continuous X-ray spectrum is extremely weak; it *can* be photographed, but with the absorption of the foil, the exposures would be prohibitive. Therefore Sanner[14] and Skinner and Johnston[12] substituted the dense line spectrum produced by a heavy condenser spark between metal electrodes in a high vacuum. The experimental set up as used by the latter authors is shown in figure 5. They also developed a special method of photometry to correct for the irregularity of intensity of the original spark spectrum and so obtained the actual variation of the absorption coefficient with wave-length. The second difficulty is to find a base on which to evaporate the absorbing foils. Since

all materials absorb very heavily in this spectrum region, the best base seems to be celluloid, on account of the fact that it can be prepared in very thin sheets, of the order of 10^{-6} cm. thick. A method of evaporating metal foils (about 10^{-5} cm. thick) on such a base was also worked out[12].

§ 3. DESCRIPTION OF RESULTS

We are now left with the task of illustrating as simply as possible the results which are being obtained from soft X-ray spectra. The results are by no means complete, and, as in all detailed spectroscopy, they are not free from the complication inherent in the material which is built up from more or less complex atoms, with the added difficulty of a more or less complex crystal structure. As we have stated, the general problem is to find out as much as possible about the way in which the energy levels are distributed in a crystal, in other words, to determine the density curve for the levels, which we have called the $N(E)$ curve.

We have seen in § 1 that corresponding to a given X-ray level there exists (1) a valence-emission band which gives us information about the density curve for the filled levels, and (2) an absorption-structure band which gives us information about the density curve for the levels which are empty in the normal crystal. The measurements consist: (i) of the intensity distribution $I(E)$, measured in numbers of quanta,* as a function of the emitted frequency, and so of the energy E, and (ii) of the absorption coefficient $\mu(E)$ of a layer of the material, as a function of the frequency of the absorbed radiation, and so of the energy E. We may choose E most conveniently such that $E = h(\nu - \nu_0)$, where ν_0 is the smallest emitted frequency of the emission band in question. Thus $E = 0$ for the lowest level of the band of filled levels. We may then write down equations connecting $I(E)$ and $\mu(E)$ with $N(E)$:

$$I(E) \propto \nu^3 f(E) N(E) \quad E \leqslant E_m, \qquad \text{.....(1)}$$

$$\mu(E) \propto f(E) N(E) \quad E > E_m \qquad \text{.....(2)}$$

Here E_m is the maximum energy of the *filled* levels, and $f(E)$ is the square of the "dipole moment" for the relevant transition of energy $h\nu$. The factor ν^3 is that which always enters when the ratio of transition probabilities for emission and absorption is considered. It is clear that, owing to the presence of the factor $f(E)$, which depends on the precise wave functions of the electron in the normal and excited states and which controls the transition probabilities, $I(E)$ and $\mu(E)$ can only be expected to represent $N(E)$ *directly* in a very qualitative way. We shall return to this question in § 3 (*d*); meanwhile there are some important properties of the bands which do not depend on $f(E)$ and we shall deal first with these.

(*a*) *Metal absorption and emission edges.* According to the theory of metals the difference between a metal and an insulator lies in the fact that in the former there are *unoccupied levels immediately above the filled valence levels*, while in the latter there are not. Possible $N(E)$ curves, based on the work of Jones[13] and Jones

* Since this corresponds most closely to what is recorded by the photographic plate.

and Mott[16], for the two cases are sketched in figure 7. It is easy to see that the property of conduction is explained in this way: in a metal, an electron can have its energy slightly raised by acceleration in an applied electric field, since there are levels in which it can exist and run through the crystal as a conduction electron; while, in the insulator, no such levels are available.

From such a model, it follows from equation (1) that for a metal the curve $I(E)$ of an emission band will fall suddenly from some value to zero as E is increased beyond the value E_m at which the surface of the filled levels occurs. Conversely, an extra absorption will set in suddenly when the absorbed quantum has sufficient

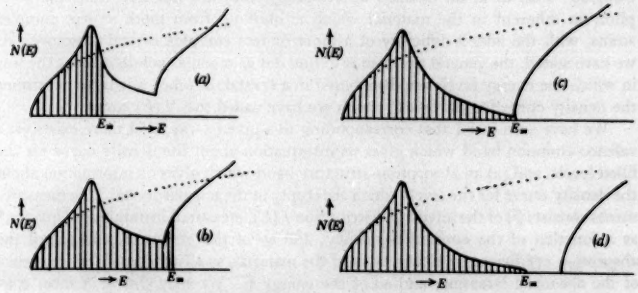


Figure 7. Illustrative $N(E)$ curves: (*a*) 1-electron metal; (*b*) 2-electron metal; (*c*) semi-conductor; (*d*) insulator.

energy to carry an electron from an inner shell into a level of energy just greater than E_m. These sudden jumps in $I(E)$ or in the absorption coefficient $\mu(E)$ are *the emission and absorption edges*;* for a given metal, they should obviously occur at the same wave-length (see figures 2 and 3).

We have now considerable material to show that this view is correct. In the work of Siegbahn and Magnusson[17] and of O'Bryan and Skinner[13] it was found that *all the metals of the first two groups of the periodic table showed such sharp emission edges*, which had first been observed by Osgood[18] in the case of Al. A new photometer curve obtained by the author for the L_2 band† of Na metal is given as the simplest example (figure 8). It will be seen that the metal radiation has a clear emission edge whose breadth is actually less than 0·1 v. To illustrate the case of an insulator, since it has been shown, § 3 (*f*)), that the case of compounds is not less easily interpreted, it is better to keep to examples of element insulators. Only two such substances have been studied, the diamond (carbon K radiation)

* Absorption edges are a commonplace in the language of hard x rays. If the resolution is only a few volts (of the order of magnitude of the energy gaps in insulators) then the distinction between conductor and insulator from our point of view is lost.

† For simplicity, we use L_1, L_2 and L_3 in place of the more usual L_I, L_{II} and L_{III}.

investigated by Siegbahn and Magnusson[19] and Broili, Glocker and Kiessig[20], and sulphur (L_{23} radiation) investigated by Siegbahn and Magnusson[17]. In the latter case, it is clear that there is no sharp edge, but unfortunately no photometer curve is given; in the case of the diamond, we give in figure 9 the result of Siegbahn and Magnusson[19]. It will be seen that there is a fairly sharp fall which occupies an energy range of several volts, but no edge which can be compared with that of sodium. Besides these true insulators, there are the very poor conductors, whose specific resistance is greater by a large factor than that of a metal. Such are B, C (graphite) and Si. None of these shows an edge of any appreciable magnitude. The case of graphite is also shown in figure 9 and incidentally illustrates the difference in the soft X-rays bands of two alotropic forms of an element. Since, as we shall subsequently see, the L_3 bands of the metals of the second group show a more

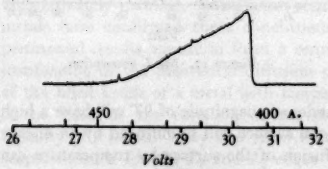


Figure 8. Na L radiation.

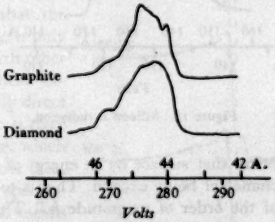


Figure 9. Carbon K radiation.

pronounced edge than the K bands of the first group metals, the case of Si is especially interesting. The photometer curve of O'Bryan and Skinner[13] is shown in figure 10.

Experiments on absorption edges in the soft X-ray region are rather more scarce. A sharp L_3 edge for Al metal has been observed by Sanner[14].* O'Bryan[21]† observed sharp edges in Li K and Mg L_{23} absorption, and his results were confirmed in more detailed work by Skinner and Johnston[12], who determined the actual absorption coefficient of the metals as a function of wave-length. A sharp edge was also found in the M_{23} absorption of Ni. Their result for Mg is given in figure 11, in which $\Delta\mu$ represents the actual variation of the absorption coefficient. The L_2 and L_3 edges can be separately distinguished, and the edge diffuseness in this case is certainly less than $\frac{1}{4}$ v. In all examples for which we have the necessary data, *the absorption edge of a metal agrees exactly in wave-length with the corresponding emission edge.*

There are no absorption data for element insulators. The nearest approach is

* The striking result of Sanner on the M_{45} absorption of Se requires the corresponding emission band and possibly a higher dispersion for its elucidation.

† In this paper there are also data of the K and L_{23} absorption of Li, Na and Mg combined as halides. There is a line absorption, which emphasizes the unusual characteristics of the halides as solids which were referred to on p. 257.

the case of graphite, whose K absorption has been studied by Magnusson[22]. As far as one can tell, the absorption begins just at the extreme short wave-length limit of the emission band, as would be expected.

(*b*) *Temperature effect on a metal emission edge.* According to the theory of metals, the surface of the filled conduction-electron levels in a metal should only be sharp at the absolute zero of temperature. At a temperature T, electrons lying

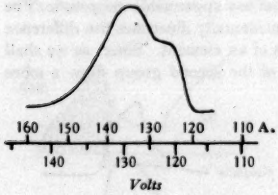


Figure 10. Silicon L radiation.

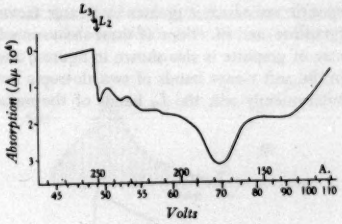


Figure 11. Mg L absorption.

below that surface by an energy of the order of magnitude of kT will have a high chance of being excited. That is to say, the surface will be diffused by an energy of the order of magnitude kT. The diffusion of the surface by temperature can be easily calculated. The result is that, at a temperature T, the number of levels which are *occupied* is given by the formula*

$$n(E) = N_0/(1 + e^{\delta E/kT}). \qquad \ldots\ldots(3)$$

In this formula $N_0 = N(E_m)$, where E_m is the energy of the free surface at the absolute zero of temperature and $N(E)$ is assumed constant in a small energy-range around this value E_m. δE is the energy (positive or negative), measured from the value E_m.

By running an anticathode at various temperatures, Skinner[24] has been able to verify this formula in the case of Al metal.† Figure 12 shows, on a large scale, the Al L_3 and L_2 edges, obtained at temperatures from that of liquid oxygen upwards. The diffusion of the edge is made up of various factors; there are (*a*) the slit-width of the spectrograph, (*b*) the width of the underlying X-ray level and (*c*) the temperature of the metal. In figure 12 the results for three temperatures are given. The effect of the factor (*c*) is clearly seen, since the factors (*a*) and (*b*) are constant. The edge breadth can be measured; subtracting the value of (*a*) as determined by a line spectrum, the results are given in table 1. Measurements on the edge

* See e.g. Mott and Jones[23], p. 175.

† Hautot and Serpe[25] report a smoothing out of characteristic features of the boron K band at a high temperature (1500° K.). Since this band has no sharp edge, this is a different effect from the above; it can perhaps be interpreted as due to lattice irregularity introduced by the heat motion of the atoms.

Table 1

Temp. °K.	Corrected edge breadth v.	6 kT v.
110	0·06 ± 0·03	0·057
300	0·17 ± 0·03	0·156
680	0·44 ± 0·05	0·354

breadth correspond to about 95 per cent of the total, assuming that one could measure from $\delta E = -\infty$ to $\delta E = +\infty$. According to equation (3), this should correspond to about $6kT$, and the values of this quantity are appended. It will be seen that the agreement is quite good if one bears in mind the fact that the constancy of $N(E)$ in the neighbourhood of E_m is only approximately justified. Subsequent results with other metals have confirmed these conclusions. These experimental results appear to form a remarkably direct justification of the theoretical diffusion of the surface of the filled levels of a metal with temperature, which is the basis of the variation of the electronic heat capacity of a metal. They also show that the spectroscopic breadth for the L_3 and L_2 edges of the metals of the second group (i.e. the width of the underlying X-ray level, which in § 1 we have defined as the inverse of the resolving power of the method) at the absolute zero of temperature is very small, being at most a few hundredths of a volt.

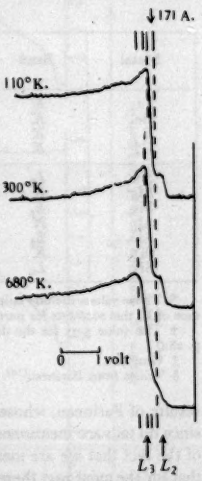


Figure 12. Al L_3 and L_2 edges, excited at various temperatures.

(*c*) *Total breadths of the emission bands.* It was pointed out by O'Bryan and Skinner[13] that the *total energy breadths* of the soft X-ray emission bands of metals of the first two groups agree fairly exactly with the values predicted by the simple Sommerfeld formula:

$$E_m = h^2/2m\,(3\rho n/8\pi)^{\frac{2}{3}}, \qquad \ldots\ldots(4)$$

where ρ is the number of atoms per c.c. and n the number of free electrons per atom.

The breadth of an emission band is not a very exact conception: it has been found that, given sufficient exposure, the bands (especially the L_{23} bands) tail off almost indefinitely towards long wave-lengths, and it has been shown that this is not a spurious instrumental effect. This tailing will be noticed in the curves to be given in figure 13. The fraction of the total radiation in the tails is very small (of the order of a few per cent of the total) and it seems possible that they are of secondary origin, such as a degeneration of the radiation by a kind of Raman effect. They

make necessary an empirical definition of the band widths; and, in the case of L_{23} bands, this may be done by a linear extrapolation of the intensity curve near the tail to cut the E axis. A similar extrapolation (as E^2) is used to define the band widths in the case of K spectra. These methods have been chosen in order to obtain definite results in a simple manner. It does not follow that the widths obtained are those required from a theoretical point of view. Though the latter can scarcely be assumed greater than the values to be given, it is quite possible (see p. 275) that they may be about 10 per cent smaller.

The results, partly from O'Bryan and Skinner[13], corrected on account of the new definitions, but mainly from new sources, are given in table 2, together with three

Table 2

Metal	Band	Measured breadth v.	Sommerfeld breadth v.	No. of electrons per atom
Li	K	$4{\cdot}1 \pm 0{\cdot}3$	4·8	1
Be	K	$14{\cdot}8 \pm 0{\cdot}5$	14·6	2
B	K	$29{\cdot}5 \pm 2$*	25·2†	3
C‡	K	30 ± 2*	21·9	4
Na	L_3	$3{\cdot}0 \pm 0{\cdot}2$	3·2	1
Mg	K	$7{\cdot}3 \pm 1$§	7·3	2
Mg	L_3	$7{\cdot}6 \pm 0{\cdot}3$		
Al	K	$12{\cdot}7 \pm 1$§	11·9	3
Al	L_3	$13{\cdot}2 \pm 0{\cdot}5$		
Si	K	18 ± 2§	12·7	4
Si	L_3	$18{\cdot}2 \pm 1$		

* These values include tails of several volts on the *short* wave-length sides of the bands. In the case of B, this accounts for part of the divergence from the value of Hautot and Serpe[23] (21·5 v.).

† The value 3·33 for the density of boron has been taken from Landolt-Börnstein (E.G. III*a*, p. 282).

‡ Graphite.

§ Values from Farineau[10].

results of Farineau, whose values, however, are obtained in a straightforward way, since no tails are mentioned in his papers. They have been slightly corrected, in view of the fact that we are measuring to the mean points of the edges. It will be seen that for the most part there is a tendency for the measured breadth to be greater than the Sommerfeld breadth, except lithium.

The Sommerfeld theory, which treats the metal as a free-electron gas and so neglects the lattice structure altogether, can of course be regarded as only a first approximation. It would give $N(E) \propto E^{\frac{1}{2}}$. It is well known (see Mott and Jones[23] for a recent discussion) that the actual $N(E)$ curves have, in the simplest possible (idealized) cases, the forms already illustrated in figure 7 in which the Sommerfeld curve is dotted. In general the curves would be more complicated,* with several peaks; but the form given is typical. Since the area under the $N(E)$ curve up to the limit of the filled levels (E_m) must be proportional to the number of

* However, the curve for the 1-electron metal probably represents fairly accurately the case of the alkalis (body-centred cubic lattice).

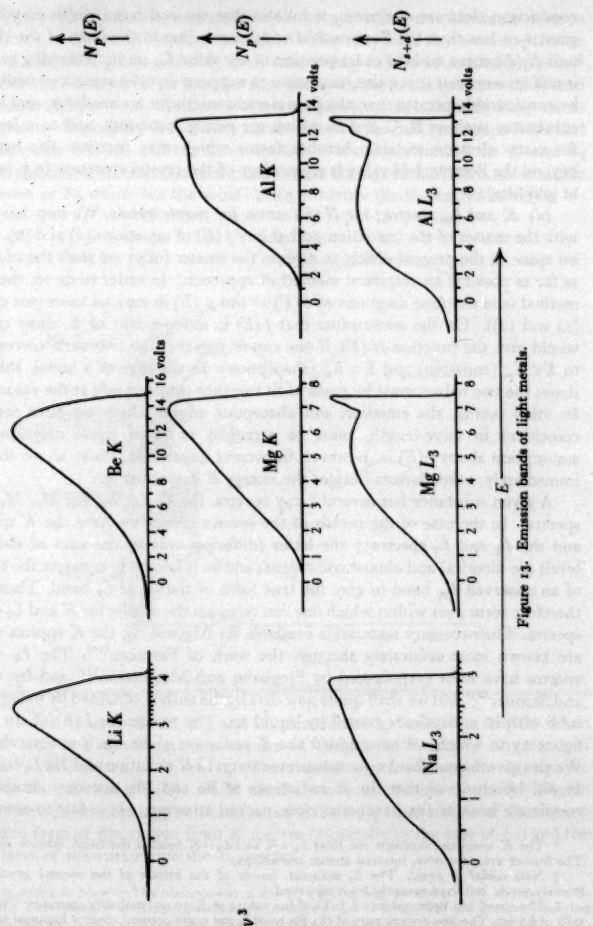


Figure 13. Emission bands of light metals.

conduction electrons per atom, it follows that the real band width may be either greater or less than the Sommerfeld width according to the form of the characteristic $N(E)$ curve and the exact position of the value E_m on it. Referring to figure 7, it will be seen that it is quite reasonable to suppose that the measured width should be considerably greater than the Sommerfeld width for an insulator, and hence for substances such as B, C and Si which are nearly insulators, and to a less degree for many electron metals. Another factor which may increase the band width beyond the Sommerfeld value is asymmetry of the crystal structure (e.g. in the case of graphite).

(*d*) *K and L_{23} spectra; the $N(E)$ curves for simple metals.* We now have to deal with the matter of the transition probability $f(E)$ of equations (1) and (2). There is no space in the present article to discuss the matter fully; we shall therefore adopt as far as possible an empirical method of approach. In order to do so, the obvious method is to examine diagrams of $I(E)/\nu^3$ and $\mu(E)$ in various cases (see equations (1) and (2)). On the assumption that $f(E)$ is independent of E, these quantities would give the function $N(E)$, if one can fit together the two parts corresponding to $E \leqslant E_m$ (emission) and $E > E_m$ (absorption). In the case of a metal, this can be done; the two halves must be made to fit together continuously at the value $E = E_m$. In other words, the emission and absorption edges, which we have seen to be coincident in wave-length, must be arranged to be of equal magnitude. The assumption about $f(E)$ is, however, in general unjustified, since, as we shall show immediately, other factors besides the energy E determine it.

A given substance has several x-ray spectra, the K, L_1, L_2, L_3, M_1, M_2, M_3, ... spectra. In the case of the metals of the second group, we have the K spectrum* and the L_2 and L_3 spectra;† the latter (differing only in the spin of the ground level) are identical and almost coincident; and so it is easy to separate the two parts of an observed L_{23} band to give the true form of the L_3 or L_2 band. These metals therefore form a set within which one can compare the results for K and L_3 emission spectra. The necessary material is available for Mg and Al, the K spectra of which are known most accurately through the work of Farineau[10]. The L_3 emission spectra have been investigated by Siegbahn and Magnusson[17] and by O'Bryan and Skinner[13], but we shall quote new data by the author, obtained by using an x-ray tube with its anticathode cooled by liquid air. The results for $I(E)/\nu^3$ are given in figure 13 to which we have added the K radiation of Be, on a comparable scale. We also give the results of new measurements on Li K radiation and Na L_3 radiation.‡ It will be observed that the K radiations of Be and Mg are very similar; these metals are both of the hexagonal close-packed structure. It is fair to assume that

* The K spectrum contains the lines $L_3 \rightarrow K$ and $L_2 \rightarrow K$ besides the band valence electron K. The former are, of course, internal atomic transitions.

† *Note added in proof.* The L_1 emission bands of the metals of the second group are extremely weak, but have recently been observed.

‡ The small but finite values of $I(E)/\nu^3$ for values of $E < 0$ are probably spurious. They are the tails of § 3 (*c*). The low-energy part of the Be band is not quite certain, since it happens to coincide with a region of selective absorption of SiO_2 in the grating. Details of all the new results will be given in a forthcoming paper. *Note added in proof.* The total breadth of the Na L_3 band has been erroneously drawn as 3·0, instead of 3·4 e.v. The scales of figure 8 are also affected by the mistake.

the two curves should be nearly identical. However, the Be curve shows an obvious fall in the intensity before the edge is reached. One may infer that a corresponding fall in the case of Mg has probably been lost, owing to lack of resolving power.* The resolution for Be is better than for Mg, partly owing to the greater volt-spread of the band and partly to the better resolving power attainable with long wavelength spectra. Both these K intensity curves have a quite different shape from that of the L_3 curve of Mg; and the same applies to a comparison of the K and L_3 curves in the case of Al (face-centred cubic). Also the L_3 curve of Na is quite unlike the K curve of Li which has the same lattice structure (body-centred cubic).†

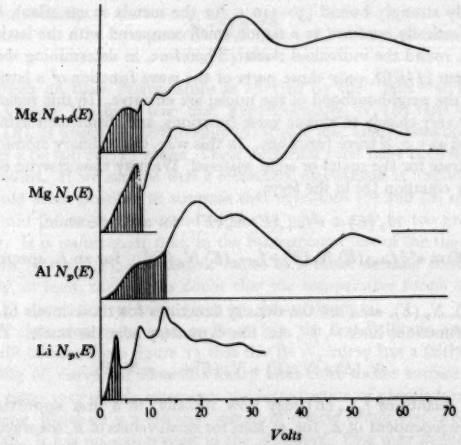


Figure 14. Experimental $N(E)$ curves.

The complete curves for Al, Mg and Li, constructed by fitting together the above emission data and the absorption results of Sandström[8] and of Skinner and Johnston[12], are shown in figure 14. For reasons which will appear shortly, the K curves are marked $N_p(E)$, while the L_3 curve is marked $N_{s+d}(E)$.† Again characteristic differences may be observed at sufficiently small values of E between the peaked form of the curves from K spectra (especially in the case of Li) and the kinked form of the curve from the L_3 spectra.

* *Note added in proof.*—The observation of the L_1 band of Mg has shown this inference to be erroneous. Theoretically and experimentally it resembles the K band, but the greater resolving power enables one to say definitely that the edge occurs at the point of maximum intensity.

† *Note added in proof.*—Owing to the remark regarding the L_1 band of Mg in the previous footnote, the dotted portion of the Mg curve is not correct. The curve should rise at the edge, in a way similar to the curve for Al.

These differences must be ascribed to the transition probability factor $f(E)$, equation (1).

This conclusion should not be considered as surprising. If we consider, for a moment, the analogy of atomic spectroscopy, it is clear that, owing to the *l selection rule*, we should determine a quite different set of levels for an atom from transitions in which the excited state of the atom was an s state, from those determined from transitions in which the excited state was a p state. The parallel with our K ($1s$) and L_3 ($2p$) spectra is obvious.

It is important to recognize that in the x-ray case we are dealing with transitions which are *very similar to atomic transitions*•; for the K or L_{23} shells, being comparatively strongly bound (30–110 v. for the metals in question), have wave-functions practically confined to a region small compared with the lattice constant of the metal, round the individual nuclei. Therefore, in determining the transition dipole moment $[f(E)]^{\frac{1}{2}}$, only those parts of the wave function of a lattice electron which lie in the neighbourhood of the nuclei are effective. In this region they will approximate very closely to atomic wave functions, and, from their symmetry, can be designated as s, p, d wave functions. In this way, the ordinary atomic l selection rule will operate for the metal or solid material. We may thus rewrite equation (1) and similarly equation (2) in the form,

$$I_k(E) \propto \nu^3 f_{p\to s}(E)\, N_p(E) \quad \text{for a } K \text{ spectrum,} \qquad \ldots\ldots(5)$$

$$I_{L_3}(E) \propto \nu^3 [f_{s\to p}(E)\, N_s(E) + f_{d\to p}(E)\, N_d(E)], \quad \text{for an } L_3 \text{ spectrum,} \qquad \ldots\ldots(6)$$

where $N_s(E)$, $N_p(E)$, etc., are the density functions for those levels of the lattice whose wave functions have s-, p-, etc. like symmetry near the nuclei. Thus:

$$N_s(E) + N_p(E) + N_d(E) + \ldots = N(E).$$

The various quantities $f_{s\to p}(E)$ may now actually, to a first approximation, be regarded as independent of E; for, at least for small values of E, the wave functions of the corresponding states may be expected, *in the neighbourhood of the nuclei*, to depend very little on E. We may therefore regard the curves which we have given in figures 13 and 14 as representing either (i) in the case of a K spectrum, $N_p(E)$ or (ii) in the case of an L_3 spectrum, $N_{s+d}(E)$; or more strictly, $[N_s(E) + \alpha N_d(E)]$ where $\alpha = f_{d\to p}/f_{s\to p}$. Of course, since a variation with E probably still remains in the functions $f_{s\to p}(E)$, the N_p and N_{s+d} curves given in figures 13 and 14 may be somewhat distorted. But any such error should be a smooth function of E, and there is little doubt that such curves as those of figures 13 and 14, which represent the experimental material in its most reduced form, give accurately the essential characteristics of the $N_p(E)$ and $N_{s+d}(E)$ curves.

It would take us too far afield to discuss theoretically the interpretation of the

• This point was neglected by Houston[26], who has the credit of being the first to use the data of soft x-ray emission spectra to obtain information about the electron levels in a solid. It was first pointed out by Jones, Mott and Skinner[27]. See also Farineau[10].

characteristic $N_p(E)$ and $N_{s+d}(E)$ curves. Reference must be made to the paper of Jones, Mott and Skinner[27].* The main points appear to be the following:

(1) It is a striking fact that e.g. Mg has K and L_3 emission bands of equal width. This means that, in spite of the fact that the ground state of the Mg atom is $3s$, throughout the spread of the filled levels of the metal, s and p states (in the above sense) are intermingled. Since, in general, the energy spread of the filled states is of the same order of magnitude as the ionization potential of the free atom, this admixture of levels of different types is not difficult to understand.

(2) At the value $E=0$, we should have as elementary and sure deductions from the Bloch theory of metals, independent even of our above treatment of the transition probabilities:

$$N_s(E) \propto E^{\frac{1}{2}}, \qquad \text{.....(7)}$$

$$N_p(E) \propto E^{\frac{3}{2}}. \qquad \text{.....(8)}$$

An exact check on these relationships is affected by the "tails" of the emission bands (see § 3 (*c*)). We have seen that, experimentally, $N_p(E)$ rises as about E^2; while $N_{s+d}(E)$, to which the relationship (7) should also apply, rises almost linearly with E, over a certain energy range. These empirical rules have been used to define the band widths. If we assume that a secondary mechanism is responsible for the tails, it would seem possible to suppose that equations (7) and (8) may be valid. If so, it would follow that the band widths of table 2 may be too great by about 10 per cent. It is unfortunate that, in the fundamental test of the theory provided by equations (7) and (8), one cannot come to a more definite conclusion. But qualitatively, at least, there is no doubt that the comparative forms of the $N_p(E)$ and $N_{s+d}(E)$ curves are correctly given by these relations.

(3) In a 2-electron metal such as Be or Mg, the first Brillouin zone† is nearly filled. It will be seen from figure 13 that the Be N_p curve has a fairly sharp peak, while the Mg N_p curve has instead a sharp kink close to the surface of the filled levels. The latter very clearly represents the introduction of s levels from the second Brillouin zone, which in a 2-electron metal, of course, must just overlap the first (see figure 7). It has no counterpart in the N_p curve. One may deduce that, if we divide the $N(E)$ curve into contributions ${}_1N(E)$ and ${}_2N(E)$ from the separate Brillouin zones, then ${}_2N_s(E)$ rises more rapidly, perhaps as $\sqrt{(E-E_m)}$, in the second zone than ${}_2N_p(E)$, a behaviour exactly analogous to what happens at the base of the first zone.‡ Such experimental facts represent the only direct evidence with regard to the overlapping of the zones in a metal.

(4) As could be established to a considerable extent in detail, if there were space in the present article, the individuality of the forms of the curves for various metals is determined almost entirely by the crystal structure. The characteristic features of the $N(E)$ curves are associated with those lattice planes which give a Bragg reflection of electrons. But in order to make precise calculations one must

* See also Jones and Mott[16] and Farineau[10].

† For explanation, see e.g. Mott and Jones[23].

‡ It is assumed above that $N_p(E)$ for Mg is the same as $N_p(E)$ for Be. In a footnote to p. 273, we have seen that this is not exactly the case. Nevertheless, the conclusion is unaffected.

proceed according to the method of Wigner and Seitz[28] in determining the wave functions of the electrons in a lattice, using as basis the corresponding atomic wave functions. Unfortunately, this is in general laborious, and no calculations of the $N_s(E)$ and $N_p(E)$ curves for the light metals have so far been given.*

Any discussion of the $N(E)$ curves would be quite incomplete without some reference to the very interesting work on the so-called fine structure of absorption edges in the region of fairly hard X rays (1 to 2 A.), which has been carried out by Coster and his school. The characteristics of the work from the present point of view are:

(1) The results can be used in conjunction with equation (2) to determine the $N(E)$ curves in the way we have described for soft X-ray absorption; in particular, there seems no reason to suppose that equations (5) and (6) should be inapplicable.

(2) The resolving power is considerably less than in the case of soft X-ray absorption; in actual cases the diffusion appears to be at least 5 or 10 V.

(3) The energy range covered is much greater than with soft X rays. Fluctuations of the $N(E)$ curve can be traced over a range of 300 or 400 V.

(4) The main factor in determining the form of the curves is the crystal structure.

As illustrating (4), it has been shown by Veldkamp[30] that the K absorptions of Cu, Ni and Ca (all face-centred cubic metals) show exactly the same structure when the units of the energy scale are taken inversely proportional to the square of the lattice constant. Veldkamp[31] has also shown that there is a similarity between the K absorption of Fe and the L_3 absorption of W (body-centred cubic metals). The absorption curves, from[32] and[31], are shown in figure 15, the energy scale units being adjusted as above. It will be seen that the curves cannot be considered as identical even at a large distance from the edges, though they show an undoubted resemblance.† On our views, it would appear that an s absorption (K or L_1) cannot be identical with a p absorption (L_3) *in detail*, even at high energies. But lack of resolving power (and possibly a lack of perfection of the lattice) certainly make it impossible for such detail to be observed; and since, at high energies, levels of all types (s, p, d, etc.) will be thoroughly intermingled in the constitution of $N(E)$, it seems reasonable to suppose that what the experiments give is, at sufficiently high energies, a kind of average of the $N(E)$ curve itself.

This would appear to be the basis of Kronig's theory[33], in which the influence of the ground level is not taken into account. He considers the motion of a free electron in a lattice, and by determining those planes of the lattice which give Bragg reflection of electrons, he is able to find the actual energy values at which maxima and minima of the $N(E)$ curve should occur. These, at high energies, are very numerous, but, taking a kind of average, a satisfactory agreement with the

* Millman[29], however, has given the results of certain calculations for Li metal. The peak of the emission band agrees with what these calculations give.

† The attempt of Veldkamp[31] to establish a resemblance between the L_1 and L_3 absorptions of a given metal (e.g. W) depends on an arbitrary energy shift (about 20 V.) of the edges relative to one another, which would be difficult to understand. In any case, the L_1 absorptions are very weak.

observed curves for different types of lattice is obtained. In the particular case of metals with similar crystal structures, the free-electron model leads easily to the result that, in comparing the absorptions, we should take energy units inversely proportional to the square of the lattice constant, as is found experimentally. These results form a remarkable justification of this model of a metal in the region of high energies.

In figure 15 a strong absorption will be noticed in the case of W immediately to the high energy side of the edge. This is a characteristic effect which is found

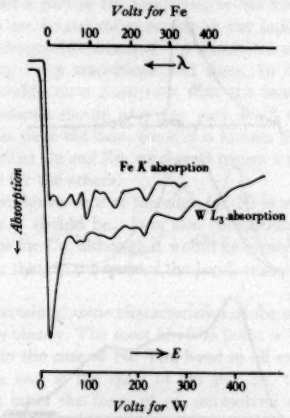


Figure 15.

with metals of the Pd and Pt groups. The absorption is obviously due to a band of vacant *d* levels, atomic in character, such as is known to exist for metals of the transition groups; see § 3 (*e*). But this absorption only occurs in cases where transitions to *d* levels are allowed by the optical *l* selection rule. In the case of W, the transition $2p\ (L_3) \rightarrow 5d$ is allowed, while in the case of Fe the transition $2s\ (K) \rightarrow 3d$ is forbidden. So Fe does not show this absorption, although the vacant $3d$ levels are known to exist.

(*e*) *Emission bands from Fe-group metals; satellites.* In the case of the *K* spectra of the second group metals, besides the main emission band, there are two "satellite" bands to be found on its short wave-length side[9]. The same is true of the *L* spectra of the Fe group metals. As is well known in X-ray spectroscopy, these satellites are due to a double ionization of the emitting atom; in certain cases this can occur, as shown by Coster and Kronig[34], by a secondary Auger process, even though it was initially ionized only once (e.g. when the electron ejected in the X-ray tube is an L_1 electron, $L_2 \rightarrow L_1$ or $L_3 \rightarrow L_1$ with the ejection of an outer

electron gives a satellite to an L_2 or L_3 band).* For satellite emission, there is the extra condition, not mentioned by Coster and Kronig, that the second ionization must *not* occur in a conduction (or valence) electron group; for, in this case, it would pass from the excited atom over to a neighbour before the radiation could be emitted. Thus the L_{23} spectra of the second group metals have no satellites.

But, in the case of the L_{23} spectra of the Fe group metals, the second ionization can occur in one of the M shells.† Therefore satellites can occur. It is rather difficult in general to distinguish between these satellites and kinks of the main

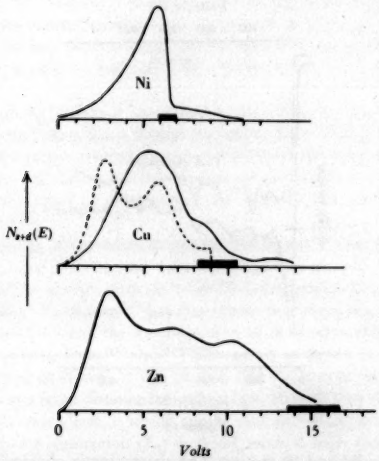


Figure 16. L_3 Emission bands.

band. The only method available is to know at what wave-length the absorption edge of the metal occurs; then all emission to the short wave-length side of this absorption edge can obviously be set aside as satellite emission.

The L_{23} emission bands of the metals Fe to Ge have been photographed by Saur[5], by Gwinner[6] and by Farineau[10], and the results agree fairly well where they overlap.

The bands of the first authors are very similar in shape to Farineau's but have a somewhat greater energy spread. Since Farineau's dispersion was considerably larger, this extra width may be an instrumental effect. Farineau's results for Ni, Cu and Zn are shown in figure 16. They have been confirmed by the work of Skinner

* The high probability of this Auger process also explains the weakness of the L_1 spectra of the second group metals.

† If an Auger process is involved, energetic considerations require that it should be in the M_{45} (3*d*) shell.

and Johnston[35] on the M_{23} emission bands of these metals. We have indicated on the diagram the energy regions covered by the corresponding apparent absorption edge breadths. In the case of Zn, the L_3 absorption curve is due to Sandström[8]; in the cases of Cu and Ni, the data are from the M_3 absorptions[12] and emission bands[35]; the results have simply been transferred to the L_3 curves.*

The results are of considerable interest since, in these cases, we are dealing, not only with transitions of the type $4s$ (conduction) $\rightarrow 2p$, but also with transitions $3d \rightarrow 2p$. We can thus obtain information about *the structure of the* $3d$ *group of levels* which play so important a part in the properties of the Fe metals. These are not conduction levels, but are bound more or less to the individual atoms. In fact, since at any rate for hydrogen-like atoms, $d \rightarrow p$ transitions are much more probable than the corresponding $s \rightarrow p$ transitions and since, in the relevant cases, the d electrons are considerably more numerous than the conduction electrons, it is doubtful if $s \rightarrow p$ transitions should play any part at all in the structure of the observed bands. If this were the case, since it is known that the $3d$ band is only partly full in Ni, but full in Cu and Zn, we should expect a sharp emission edge for the first metal, but not for the others.

It will be seen from figure 16 that the edge for Ni is much more pronounced than for Cu or Zn, as it should be. It is also considerably sharper; but one is tempted to infer an edge for Cu, although it would be a very diffuse one.† If this is correct, it would follow that, in this part of the band, transitions (conduction) $\rightarrow L_3$ cannot be neglected.

In spite of this uncertainty, some characteristics of the structure of the $3d$ band of levels stand out very clearly. The most obvious point is the effect of the incompleteness of the band in the case of Ni. The band in all cases occupies an energy range of 5 v. or more, and in the cases of Cu and Zn, there are clear signs of structure. For Cu, it takes the form of an unresolved doublet. The fact that Gwinner has found the separation between the components to be clearer for the alloy CuBe than for pure Cu shows that this lack of resolution cannot be altogether due to the experimental method. For Zn the structure is probably that of a triplet.‡ On the other hand, Ni, somewhat surprisingly, shows no signs of any structure at all, and Co and Fe show at most only traces, found by Gwinner but not by Farineau.

In the case of Cu, we can compare the result with the calculation of Slater[37, 38], using the Wigner-Seitz method. His $N(E)$ curve for Cu (conduction levels as well as $3d$ levels) is dotted in figure 16. It will be seen that, in spite of a general similarity, the agreement in detail is not particularly good. All the results seem in definite disagreement with the results of a more general calculation of Jones and Mott[16], using the approximation of tight binding; the validity of which is, indeed, made doubtful by the large energy-spread found for the $3d$ bands.

* Sandström's result for Cu L_2 absorption has to be disregarded. It would place the absorption edge almost at the maximum of the emission curve, in complete disagreement with the above.

† This edge may perhaps be an illusion since the apparent absorption edge breadth may be greater than the true breadth, and so the marked regions in figure 16 may stretch too far towards high energies.

‡ Bethe[36] was the first to show that a cubic field should split the d band into two parts. Cu has a cubic crystal structure, but Zn is hexagonal.

Finally, it was pointed out by Skinner and Johnston[35] that the observed band-width for Ni corresponds to about 0·6 conduction electrons per atom (see equation (4)). This is in good agreement with the result of a calculation by Mott and Jones[23] from the saturation magnetic intensity. But it would seem that the result may be fortuitous because, as we have seen, the conduction electrons are scarcely likely to be involved, to any considerable extent, in the observed band. The significance of the numerical agreement may therefore be that the filled conduction levels and the *d* levels occupy about the same energy range.

(*f*) *Alloys and chemical compounds.* Very little systematic work has been published on these lines. The simplest case is that of a solid solution of one metal in another. It was found first by Yoshida[39] that an alloy of a few per cent of Al

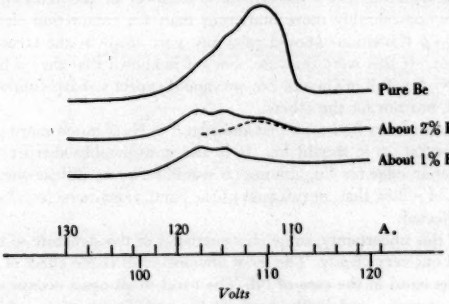


Figure 17. Be *K* band from Be-Al.

in Cu gives a result for the *K* radiation of Al very different from that from pure Al, or indeed from any metal. The main characteristic is that there is no emission edge at all. Similar results were obtained by Skinner and Johnston[40] for dilute Al-Cu, Be-Cu, and Be-Al alloys. Their result for Be-Al is shown in figure 17.* A formal explanation was also given. Briefly, this is that when, to take a simplified case, we have one impurity atom present in the material investigated, the emission spectra only give the energy distribution of the levels in the immediate neighbourhood of the impurity, and this is certainly not characteristic of the metal as a whole. The absence of the emission edge leads to the conclusion that the impurity atom does not take part in the conductivity of the alloy.

In the case of the intermetallic compounds Mg_2Al_3 and Mg_3Al_2, Farineau[10] found *K* bands of identical form for Mg and for Al somewhat differing for the two alloys, and different from the bands for the individual metals.

The *K* radiation of carbon from various carbides has been investigated by Broili, Glocker and Kiessig[20], and by Siegbahn and Magnusson[17]. In the cases

* Inflections near 117 and 115 A. are due to selective absorption by the grating.

of substances such as SiC and TiC, the detailed shape of the curves are quite different from that obtained from graphite or diamond, but the wave-length range covered is roughly the same. The same is true for the K radiation of boron emitted from various covalent compounds, as shown by Gwinner and Kiessig[41].

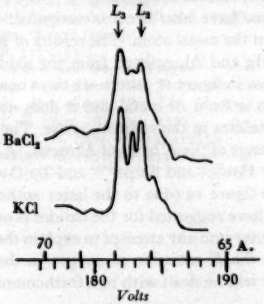


Figure 18. Chlorine L_{23} radiation.

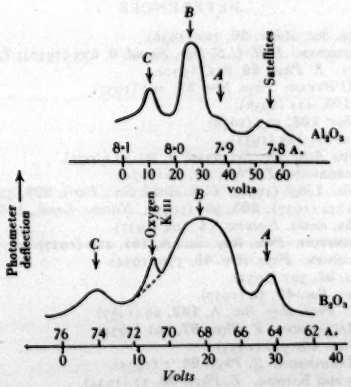


Figure 19. K emission bands from metal atoms in oxides.

Not very much systematic work has been published dealing with ionic compounds. The halides are the simplest case. Siegbahn and Magnusson[42] have found that the main parts of the L_2 and L_3 radiations of Cl from various chlorides each consists of one fairly sharp line accompanied by a fainter companion which may perhaps be a satellite (see figure 18). This proves that the $3s$ levels of Cl in a

chloride are comparatively sharp, and, from the M_{45} spectra of Br in RbBr obtained by these authors[42], it appears that the same applies to the $5p$ levels of Br. It would thus seem possible to suppose (at least as a first approximation), that the levels of the halides, especially of the alkali halides, differ only from the levels of the corresponding negative ion by a certain broadening.

A few cases of oxides have also been investigated. In these, the radiation photographed is that from the metal atom. The results of Karlsson and Siegbahn[9] on the K radiations of Mg and Al, emitted from the oxides, are probably typical. The case of Al_2O_3 is shown in figure 19; there are *three* bands (A, B and C) in place of *one* when the emission is from Al metal and it does not seem at all possible to regard any of these as satellites in the ordinary sense. The band B covers approximately the wave-length range of the K band of Al metal. A result on the K radiation of B in B_2O_3 obtained by Hautot and Serpe[25] and by Gwinner and Kiessig[41] is very similar, as shown in figure 19 (due to the latter authors). It will be seen that a simple view such as we have suggested for the halides is out of the question for the oxides. We cannot here enter into any attempt to explain these bands; it is obviously essential to have data on the K radiation of oxygen in the various chemical combinations, and the matter will be dealt with in a forthcoming paper of O'Bryan and Skinner.

REFERENCES

(1) O'Bryan. *J. Opt. Soc. Amer.* **26**, 122 (1936).
(2) Mohler and Boeckner. *Bull. U.S. Bur. Stand.* **6**, 673 (1930); **7**, 751 (1930).
(3) Hilsch and Pohl. *Z. Phys.* **59**, 812 (1930).
(4) Schneider and O'Bryan. *Phys. Rev.* **51**, 293 (1937).
(5) Saur. *Z. Phys.* **103**, 421 (1936).
(6) Gwinner. *Z. Phys.* **108**, 523 (1938).
(7) Johann. *Z. Phys.* **61**, 185 (1931).
(8) Sandström. *Nova Acta Soc. Sci. upsal.* **9**, no. 11 (1935).
(9) Karlsson and Siegbahn. *Z. Phys.* **88**, 76 (1934).
(10) Farineau. Thesis. Liège (1938). *C.R. Acad. Sci., Paris*, **203**, 540 (1936); **204**, 1108 (1937); **204**, 1242 (1937); **205**, 365 (1937). *Nature, Lond.*, **140**, 508 (1937).
(11) Siegbahn. *Ergebn. exakt. Naturw.* **16**, 104 (1937).
(12) Skinner and Johnston. *Proc. Roy. Soc.* A, **161**, 420 (1937).
(13) O'Bryan and Skinner. *Phys. Rev.* **45**, 370 (1934).
(14) Sanner. *Z. Phys.* **94**, 523 (1935).
(15) Jones. *Proc. Phys. Soc.* **49**, 50 (1937).
(16) Jones and Mott. *Proc. Roy. Soc.* A, **162**, 49 (1937).
(17) Siegbahn and Magnusson. *Z. Phys.* **87**, 291 (1934).
(18) Osgood. *Phys. Rev.* **44**, 517 (1933).
(19) Siegbahn and Magnusson. *Z. Phys.* **96**, 1 (1935).
(20) Broili, Glocker and Kiessig. *Z. Phys.* **92**, 27 (1934).
(21) O'Bryan. *Phys. Rev.* **49**, 944 (1936).
(22) Magnusson. *Nova Acta Soc. Sci. upsal.* **11**, no. 3 (1938).
(23) Mott and Jones. *Theory of the Properties of Metals and Alloys* (1936).
(24) Skinner. *Nature, Lond.*, **142**, 432 (1938).
(25) Hautot and Serpe. *J. Phys. Radium*, **8**, 175 (1937).
(26) Houston. *Phys. Rev.* **38**, 1791 (1931).
(27) Jones, Mott and Skinner. *Phys. Rev.* **45**, 370 (1934).
(28) Wigner and Seitz. *Phys. Rev.* **43**, 804 (1933).

(29) MILLMAN. *Phys. Rev.* **47**, 286 (1935).
(30) VELDKAMP. *Z. Phys.* **82**, 776 (1933).
(31) VELDKAMP. *Physica*, **2**, 25 (1935).
(32) COSTER and VELDKAMP. *Z. Phys.* **74**, 191 (1932).
(33) KRONIG. *Hand. d. Phys.* **24/2**, 295 (1933).
(34) COSTER and KRONIG. *Physica*, **2**, 13 (1935).
(35) SKINNER and JOHNSTON. *Nature, Lond.*, **140**, 508 (1937).
(36) BETHE. *Ann. Phys., Lpz.*, **3**, 137 (1929).
(37) SLATER. *Phys. Rev.* **49**, 537 (1936).
(38) RUDBERG and SLATER. *Phys. Rev.* **50**, 150 (1936).
(39) YOSHIDA. *Sci. Pap. Inst. Phys. Chem. Res., Tokyo*, no. 621, 243 (1936).
(40) SKINNER and JOHNSTON. *Proc. Camb. Phil. Soc.* **34**, 109 (1937).
(41) GWINNER and KIESSIG. *Z. Phys.* **107**, 449 (1937).
(42) SIEGBAHN and MAGNUSSON. *Z. Phys.* **95**, 133 (1935).
(43) SIEGBAHN and MAGNUSSON. *Z. Phys.* **88**, 559 (1934).
(44) PRINS. *Physica*, **2**, 31 (1935).

THE USE OF X RAYS AND γ RAYS IN MEDICINE

W. V. MAYNEORD, D.Sc., F.Inst.P.
Physics Department, The Royal Cancer Hospital (Free), London, S.W. 3

§ 1. INTRODUCTION

In attempting to describe recent progress in the use of X rays and γ rays in medicine we must limit the field of discussion to the therapeutic aspects of these radiations as opposed to their diagnostic applications. Although the latter constitute the major interest of most medical men using X rays, the subject may be left aside with little regret by physicists, for recent progress in the field of diagnosis, though rapid and triumphant, is of specialized medical or technical interest. On the other hand it may reasonably be claimed that progress of the last few years in radiation therapy has been largely due to the partial solution of certain physical problems.

§ 2. UNITS OF DOSE

The subject most likely to be discussed when hospital physicists meet is dosage, and a great deal of the work in medical radiology is directly or indirectly linked with this aspect of physics in medicine. It is of the utmost importance that the quantity of radiation delivered to a patient should be precisely defined and measured, so that it is not surprising that almost all the physical effects of X rays should have been at some time made the basis of a system of clinical measurement. We may quote as examples the photographic effects, the changes of colour of barium platinocyanide pastilles, the luminescence of a zinc sulphide screen, or the resistance of a selenium cell. This tendency still persists as may be seen in a recent reference[1] to measurement of dose by the use of a Weston photoelectric cell. As the relative effects on all these types of measuring apparatus of the same amount of energy at different wave-lengths are different, the need for some standard system is only too obvious. From a medical point of view the intensity of a beam of X rays or γ rays is of little interest since it frequently happens that only a small fraction of the energy falling on the mass of tissues considered is absorbed, and this fraction will clearly be dependent upon the penetration of the primary radiation. It would be expected that a measurement of energy absorption would constitute the most satisfactory type of unit of dose, since the fundamental changes brought about by X rays are measured, or at least expressed, in terms of energy changes in the atoms or molecules affected.

The unit of dose should be intimately related to the absolute c.g.s. system and not arbitrarily dependent on one experimental arrangement, while the ideal method of stating dose would enable us, at least theoretically, to express the energy absorbed

per unit volume throughout the irradiated mass. Owing to the discontinuous nature of the process and materials involved, the result must in practice be a statistical mean over a large number of units of volume of varying composition. It must also be emphasized that there is not necessarily any simple relationship between mean energy absorption and biological effect even if we restrict our attention to one wavelength only. Nevertheless energy measurements will doubtless play an important part in the more profound biochemical and biophysical theories of the future.

The amount of energy absorbed from an X-ray or γ-ray beam during a treatment is generally very small. For instance, the order of the amount of radiation administered at one time to a patient might reasonably amount to only 20,000 erg/cm^3 near the skin and this amount of radiation administered to the whole of the body would probably prove fatal though there is only sufficient energy to raise the temperature some 10^{-3} °C. Direct heat measurement in the tissues is not feasible when it is remembered that the radiation sets in progress new chemical reactions among those already existing. The latter will themselves cause an evolution of heat of a much greater order of magnitude. It may be added that attempts have been made to use an absolute energy system of dosage in radium and X-ray therapy, particularly by Murdoch and Stahel[(3)]. This system of measurement has been abandoned in favour of the ionization method now to be described.

The physical effect of X rays which most readily lends itself to accurate and convenient measurement is the ionization produced in air by the radiation, and it is a very fortunate circumstance that the mean atomic numbers of air, water and soft animal tissues are at least of the same order of magnitude, for it enables us to develop a system of dosage measurement in tissues by the observation of the saturation currents produced in air under specified conditions. That is, we replace the difficult measurement of energy absorption in a mass of tissues, by the easier measurements of ionization produced in the same mass of air.

The type of standard instrument most often selected in dosage measurements is the so-called parallel plate chamber. The beam of radiation to be measured is sent between the plates of a parallel plate condenser and ionizes the gas between the electrodes. If a sufficiently large difference of potential is maintained between the plates a saturation current is measured which represents the rate of production of ions by the beam. The beam is limited in cross-section by lead or gold stops containing apertures of accurately known sizes, simple cylindrical holes being preferred since their sizes may be precisely measured with steel gauges. The use of adequate guard plates ensures that the collecting electrode is attained only by ions formed in a specified length of path of the beam. The electrodes must be sufficiently far apart for the electrons to lose all their energy in ionization before striking the collecting electrodes, which are of course not irradiated. It is now realized that it is also important that the distance from defining aperture to measuring electrode be at least as great as the forward range of the electrons, for some of the ions collected are produced by electrons set in motion in the space between the diaphragm and this electrode. There is however a compensating loss of ionization by electrons formed in the active space but passing beyond the collector. The compensation

of the two effects is only complete if equilibrium between electrons and primary radiation has been attained at the collecting electrode, and this equilibrium clearly requires a mass of air equal to the forward range (in g./cm^2) for it to be built up. The total width of the electrode plates must also equal twice the lateral range. It is perhaps worth noting that there was at one time doubt as to the precise position at which the radiation was being measured. If the volume of the irradiated air is defined as the product of area of aperture and length of collecting electrode, then the dose is measured at the aperture of the chamber, but if it is defined by direct measurement of the limits of the beam at the collecting electrode (a method which owing to the divergence of the beam evidently gives a larger value for the cross-section) then the point of measurement is usually very close to the geometrical centre of the collecting volume.

Various means have been used for the measurement of the currents, but the Townsend compensation method is certainly one of the most suitable, as the collecting electrode may then be kept rigorously at the same potential as the earthed guard plates and no field distortions are introduced. It is obvious that the observations of ionization current must be reduced to standard conditions of temperature and pressure. The number of units of dose is the number of electrostatic units of quantity of electricity swept out per unit mass ionized. The literature dealing with the subject of standard parallel plate chambers is now quite extensive[3,4,5,6,7,8,9,10] and the conditions of measurement required for accuracy are well known. Such factors as the relative width of guard electrodes to gap or the plate spacings at different voltages have been worked out in detail for the radiations produced at potentials up to 400 kv. In an international comparison of standards[9,11] during 1931 of the apparatus of the National Physical Laboratory, the Bureau of Standards, the Reichsanstalt and a French standardizing laboratory, agreement was reached to $\pm 0{\cdot}5$ per cent up to 200 kv.

With such apparatus we may embody the theoretical conception of the "röntgen", which was defined in 1928 as "The quantity which, when the secondary electrons are fully utilized and the wall effects of the chamber avoided, produces in 1 cm^3 of atmospheric air at 0° C. and 760 mm. of mercury pressure, such a degree of conductivity that one electrostatic unit of charge is measured at saturation current". The acceptance of this international röntgen (designated by the symbol r.) represented a tremendous step forward in radiation therapy.

One of the most important developments of the last few years has been the attempt to use much higher voltages in x-ray therapy, so that it became necessary to study the suitability of the röntgen for the measurement of shorter wave-length radiations. At the same time developments of radium production and technique have made available relatively large quantities of this element which have come to be used in a manner precisely analogous to that of an x-ray tube, a beam of γ rays being directed on to the patient from outside. Developments in the low voltage x-ray field also conspired to make it a matter of considerable therapeutic importance to try and link up the dosage measurements of x rays and γ rays over a wide range of wave-lengths, for it has always been a burning question of medical radiology

whether the same dose at different wave-lengths gives the same biological effects and clinical results. Clearly the first step must be the use of a common unit of dose for the radiations.

We may anticipate by saying that at the Fifth International Congress of Radiology held during September of last year (1937) the röntgen was provisionally adopted for general use. The definition was suitably adjusted so as to make it applicable to γ rays, at the same time retaining its significance for the field previously covered.

It may be thought that this step represented an obvious decision, but in fact a very great deal of work was necessary to prepare the ground. The difficulties arise largely from that of defining γ-ray beams without introducing very large quantities of secondary radiations. The very long paths of recoil electrons produced by the γ rays in air at normal pressures make the position of measurement difficult to define.

We must now consider the ionization, not in the vicinity of parallel electrodes between which the beam of radiation passes without striking any material other than air, but in an enclosed ionization chamber with the rays passing through the walls. The ionization is then largely due to the secondary electrons emitted from those walls rather than from the air itself. The adequate discussion of the physics of such a chamber would require a volume but there appears to be one set of conditions in which the current has a relatively simple significance. If the atomic number of the walls of the chamber is the same as that of the gas inside, the ionization current is proportional to the pressure of the gas over a wide range of pressures and may be taken as a measure of the energy absorbed in the material of the walls. The fundamental theory may be found in Bragg's *Studies in Radioactivity* (1912), while its significance for therapy was developed largely by Fricke and Glasser[12]. This important condition has been re-emphasized recently in connexion with energy determinations of γ rays[13, 14]. In practice it implies that if the walls of the ionization chamber have the same atomic number as air, then the current inside should be equal to that produced in the same column of free air, that is in a parallel plate chamber. This conclusion has been tested experimentally[10] and found to be true in some conditions to a fair degree of accuracy for moderate voltage X rays. In practice we have to use mixtures of various substances, such as carbon and magnesium, to produce the air-wall material and the conditions are not quite so simple. It does appear, however, that the indications of such a chamber can be made to run parallel to those of the standard (that is, parallel plate) chamber over a wide range of wave-lengths even if the absolute values of the currents are not exactly identical. Such an air-wall chamber on the end of a flexible lead is clearly a more suitable instrument for clinical use than a delicate parallel plate chamber weighing several hundredweights. Moreover, the air-wall chamber by suitable design can be made practically independent of the direction from which the radiation comes and may therefore be used in circumstances in which the directional parallel plate type is useless. This is a very important matter, for perhaps the most important part of physical measurement in therapy

consists in the determination of the radiation at a depth in tissues, in which circumstances a large fraction of the radiation (often up to 80 per cent) is scattered from various directions to the point in question.

The conditions regarding an air-wall chamber would be expected to hold for γ rays. Indeed, they might be expected to hold more exactly for γ rays than for x rays, since for very short wave-lengths the electronic emission is almost entirely due to recoil electrons and less dependent on the atomic number than for the longer wave-lengths, while the normal dimensions of the chambers would be small compared to the range of the electrons, under which conditions the energy relations hold more exactly. With such ideas in mind attempts were made to measure γ rays in röntgens using air-wall chambers[15,16,17,18]. For therapeutic use, the constant which was required to found a rational system of dosage for γ rays is the quantity of radiation at 1 cm. distance from a point source of 1 mg. of radium surrounded by a filter of 0·5 mm. of platinum in one hour[16,19], and therefore attempts were made to measure it. This quantity is continually recurring in mathematical studies of distribution. The problem is essentially that of measuring Eve's Number[20,21], but the conditions quoted above are much more appropriate to the clinical applications. In early experiments the value of 9 r. was found[16], and from the values of Eve's number a value of 8·7 r. may be deduced[17]. More recently a number of determinations have been made[22], and the best value of the constant is regarded by the writer as $8{\cdot}3 \pm 0{\cdot}2$ r. The problem now turns on the precise measurement of content of the radium sources, the corrections for absorption in the walls and in the materials of the sources which are employed, and such geometrical questions as the exact position at which the measurement is made with a chamber of necessarily finite dimensions, or the absorption corrections and extrapolation for the finite thickness of chamber wall which must be used to build up the electronic equilibrium. At large distances from the source in normal-sized rooms scattered γ rays also play an important part in the ionization. We shall not attempt here to discuss these questions but references are given to papers in which detail may be found.

It is clearly possible to measure the same quantity with a parallel plate chamber of standard design[23], but when this was carried out[10], with a chamber which had been tested up to 400 kv., the value of the constant obtained was approximately 3·0 r. in place of the 8·3 r. previously obtained, although exactly the same measuring system was employed. Other values were obtained by other observers. The cause of the divergence was realized to be the very long paths of the recoil electrons in air, so that a parallel plate chamber must be very large to enable the equilibrium to be built up in front of the collecting electrodes and to capture all the ions set free. This has been proved experimentally[24] by recent investigations at the National Physical Laboratory, in which a parallel plate chamber of the guarded electrode type was employed with a plate spacing up to 3 m. Such a chamber yields the value of 8 r. for the constant required at small distances from a radium source, so that there is now reasonable agreement between the values by the two methods, and it appears that the röntgen may be used for the measurement of x rays or γ rays for therapeutic purposes.

This is perhaps an appropriate place to say something of the present stage of dosage measurements. We have seen that the problem has been solved by the choice of the ionization produced in air under somewhat arbitrary conditions as a measure of the biologically effective parameter, but what is the relation between röntgens recorded and the energy absorbed in the tissues? It is increasingly realized that this relationship is complex in a number of ways but we shall discuss only two of them.

It is first of all necessary to distinguish between two quite separate quantities which up to the moment we have tended to treat as identical. Consider a beam of penetrating radiation passing through a homogeneous medium in such a way that the dosage-rate (that is, the number of röntgens per unit time) does not vary appreciably in a distance small in comparison to the range of the electrons liberated. If we consider an element of volume dV in this medium, electrons are set free in dV which will lose their energy far beyond its confines. If, however, the radiation is uniform and the medium homogeneous it is easily seen that the energy lost from dV is exactly compensated by the energy gained by dV from other electrons originating outside. We shall speak of this condition as the state of compensation. In many cases of irradiation of a biological medium such as animal tissues these conditions are fulfilled only very approximately. For example the dosage-rate will often vary very rapidly from place to place, introducing geometrical factors which prevent compensation[25] (as, for instance, near an approximate point source of γ rays inserted in tissues). Alternatively the chemical composition or even density of the medium may vary from place to place so rapidly (as in different parts of a tissue or single cell) as to produce the same effect. An important set of circumstances also occurs near the surfaces of tissues such as skin when the sensitive layers of cells may be at such levels below the surface that the equilibrium electronic emission has not been attained, a problem analogous to those involved in discussions of the ionization in air-wall chambers as the wall thickness of the chamber varies. Any discontinuity of density or composition, such as the surface of a bone, will provide another example. We must then distinguish between two quantities, (*a*) the energy "conversion" in an element of volume, that is the fraction of the energy of the electromagnetic radiation which is converted into kinetic energy of electrons, and (*b*) the real energy absorption, that is, the energy abstracted from the high speed electrons in causing ionization, excitation or chemical changes during their passage through the element. When we deal with relatively low kilovoltages or focus our attention on relatively large blocks of tissue the distinction is unimportant, but for γ rays when the range of the electrons may reach some 5 mm. of tissue it is essential. In view of these and other difficulties the definition of the röntgen was revised in 1937 to the following form. "The röntgen shall be the quantity of X or γ radiation such that the associated corpuscular emission per 0·001293 g. of air produces, in air, ions carrying one electrostatic unit of quantity of electricity of either sign." The convenience and formal correctness of the change from 1 cm^3 of air at N.T.P. to 0·001293 g. of air is obvious. The definition seems to the writer to imply the measurement of energy conversion, and so the röntgen becomes essentially suitable

for use in the standardization of x- or γ-ray beams, but the energy absorption in tissues[26, 27] must be deduced from further physical investigations. Other workers prefer a somewhat different approach to the problem[28]. We must also remember that in practiee the distinction drawn is often macroscopically unimportant and it is perhaps of interest to consider briefly what order of differences in energy conversion are to be expected in different substances likely to be present in tissues. The results quoted are rough first approximations recently made[26].

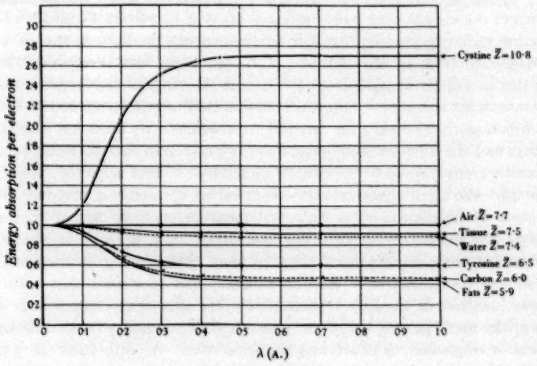


Figure 1.

If we consider the absorption of x rays in a chemical compound it is clear that the energy conversion is additive for all the atoms present and may be written

$$\mu/\rho = \mu_e \frac{[\Sigma n_1 Z_1]}{[\Sigma n_1 A_1]} N,$$

where μ is the linear absorption coefficient, ρ the density, μ_e the electronic absorption coefficient, and $\Sigma n_1 Z_1$ the sum of all the products nZ, n being the number of atoms of the type Z in the molecule, while $\Sigma n_1 A_1$ is the corresponding sum for the atomic weights A. N is the usual Avogadro's number. It is also possible to define an effective atomic number $\overline{Z}$ for the compound and so to calculate the real energy absorption (strictly "conversion") per electron or per unit mass. This may be carried out for such compounds as carbohydrates, lipoids and amino-acids. We may also try to calculate the effective atomic number of a composite tissue.

The ordinates of figure 1 show the energy conversion per electron in these materials per röntgen, compared to air, for a wide range of wave-lengths. We may remark that of course we know very little of the relative ionizations in these various materials per röntgen. It will be seen that one röntgen corresponds to only about half the energy conversion in fats that it does in air, while cystine converts some

three times as much per electron. For wave-lengths below 0·05 A. all these light materials have very similar energy conversions. We see therefore that the various materials may show appreciable differences in real energy conversion, but we must remember the complexity of the circumstances before applying the results too whole heartedly. For instance, it is unlikely in practice that the materials will be homogeneous for a sufficient region of space for the equilibrium to be built up in the neighbourhood of materials of larger effective atomic number. On the whole we may say that macroscopically the energy conversion and absorption in soft tissues would be expected to be within some 30 per cent of the energy conversion in the same mass of air, while for wave-lengths below 0·2 A. the agreement is likely to be more exact. It seems that, on the average, the effective atomic number of tissues will not be sufficiently different from that of air to cause serious divergences, but that special substances such as nucleic acids or lecithins containing phosphorus, cystine containing large quantities of sulphur or thyroxine containing iodine, will have very much greater absorption than deduced from that in air and the dose in röntgens. Bone again is a tissue obviously showing divergence. It may be noted that ionization chambers made partially of animal tissues were found some years ago[27] to deviate considerably from air-wall chambers in their response at different wave-length regions.

§ 3. HIGH-VOLTAGE X RAYS

One of the main tendencies of present day medical radiology is to increase the voltage applied to the X-ray tube while at the same time retaining such important features of normal apparatus as complete enclosure in an earthed shield, a high degree of protection against unwanted radiation or the mobility of the apparatus. It is important to bear these factors in mind as a good deal of super-voltage equipment, from the point of view of the user, appears to have reached high potentials by the sacrifice of possible accuracy and ease of adjustment of the patient to the X-ray beam. It may be asked why the desire for higher voltages? The reasons prompting the change are both physical and biological. By the use of higher voltages a greater total X-ray output is obtained which enables large doses to be given in a reasonable time. Alternatively, greater potentials make possible the use of larger focal-skin distances during treatment, which, coupled with the smaller absorption coefficient of the radiation makes for greater penetration of the beam. This is important in the treatment of deeply seated lesions as it is generally found that the limiting factor to the dose which may be given at a depth, is that which must necessarily be delivered to the overlying healthy tissues, notably the skin. Again, the decreased minimum wave-length produced at high voltages attracts those whose opinion it is that the shorter wave-lengths show greater "selective action", that is, that such short wave-length radiations are more destructive of malignant cells relatively to their action on skin or other normal healthy tissue. Any method of obtaining or increasing this difference is a matter of the greatest importance. We shall see that so far as greater penetration is concerned this by no means follows

automatically at high voltages, while the belief in selective action in the short wavelength region is being largely replaced by a tendency to think that the quantity of energy absorbed and its time spacing are the more important physical factors.

A number of methods of high voltage production are available and most of them have been applied to x-ray therapy in practice. They will not be dealt with in detail here but we may briefly mention some of them.

The direct generation of voltages up to about a million volts with a transformer has been used by Lauritsen[29] and Allibone[30], while multiplying circuits have been applied to the problem by Bouwers[31], who has designed much apparatus for medical applications as well as nuclear research. An induction coil apparatus and multiple section tube due to Coolidge[33,34] was used up to 700 kv. at the Memorial Hospital in New York City. The moving belt type of electrostatic generator has been used in a high tension unit due to Van de Graaff and his colleagues at the Huntington Memorial Hospital in Boston[35], and is impressive by its apparent simplicity and relatively low initial cost. An ingenious but complex high frequency oscillator type of high voltage generator due to Sloan[36,37] has been already described in an earlier *Report*[38] and is said to yield radiation equivalent to that from a direct generator of a million volts. Tesla high frequency currents and impulse generators have also been applied experimentally to this problem of high tension generation. For voltages up to 400 kv. an ingenious and compact unit in which the transformer and tube are immersed in a single earthed tank of oil which can be moved up and down and angulated, has been developed by the General Electric Co. of America, and forms a very convenient mobile, shockproof and well protected unit. Many of the difficulties lie in the x-ray tube itself and again a number of designs from frankly laboratory experimental models to commercial sealed off and continuously evacuated types are available. Up to 400 kv. permanently sealed tubes of standard design are used, but for voltages of the order of a million the recent tendency seems to be towards continuously evacuated or multiple section tubes in series, but no finality at all has been reached in these matters. The normal voltage still used in x-ray therapy is about 200 kv. and it is perhaps fair to say that most medical radiologists are of the opinion that higher voltage installations have still to prove their superiority in clinical practice except in a few selected sites in the body. The initial, and subsequent maintenance costs are of course very heavy at high voltages.

It may be of interest to physicists generally to give in the remainder of this report a brief account of the kind of result of the fairly intensive study of treatment methods now in progress at various research centres. We will assume that there are available methods of measuring amounts of radiation in röntgens. Stated generally, what we would like to know is the number of röntgens delivered at every point in the patient during a specified exposure. Usually nothing like this information is available, but we obtain an approximation as follows. In the physical laboratory the patient is replaced by a tank of water ("water phantom") as the absorption and scattering of x rays in water is roughly that in soft tissues. A small watertight air-wall chamber is mounted so as to be movable in vertical or horizontal planes and the beam of

X rays to be investigated is sent down into the water. Wax, various woods, rice and other materials may replace the water, but in the writer's experience are less satisfactory for complete investigations. Measurements are made throughout the mass of water along a series of vertical axes separated by 1 cm., or near the edges of the field 1–2 mm. The dosage-rates are usually stated as percentages in terms of the value at the centre of the field on the surface, and by interpolation points are found at which the percentage depth doses are convenient whole numbers. By joining all the places having the same dosage-rate, the so-called isodose curves or dose contours are obtained, which give a plan of the dose distribution in the water. That these values do correspond approximately (usually to some 10 per cent)

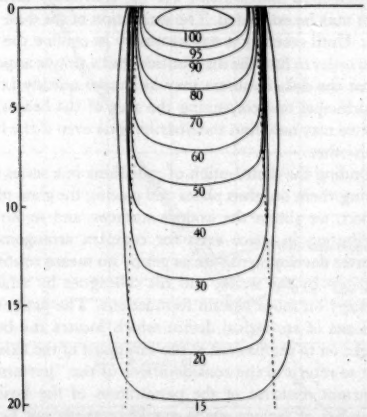


Figure 2. 400 kv. (peak); 4-mm. copper + 1-mm. aluminium filter; 80 cm. focal distance; 20 cm. × 6 cm. field; central plane parallel to short side.

to those obtaining in the body has been shown by measurements on a cadaver, in natural body cavities, or on the exit surface. We are, however, still largely ignorant of the detailed effects on the dosage distributions for X rays of the skeleton, or lung air spaces, though for γ rays we know that they have relatively little effect. A typical example of a set of dose contours is given in figure 2, which show the distribution of radiation from a 400 kv. peak, narrow X-ray beam (6 × 20 cm. rectilinear field) in a plane containing the axis of the beam and perpendicular to the long side of the field. Such a field is much used in the treatment of carcinoma of the oesophagus on to which six such fields might converge from different directions outside the body. Prior to treatment such a patient would be carefully examined and the site of the lesion estimated as accurately as possible, marks being made on the skin both back and front to mark the centres of the treatment fields which must pass through

the lesion. A cross-section of the body would then be drawn and the dose contour diagram laid in the appropriate position so that the dose contribution from each field at each point could be found. By the help of a specially designed electric adding machine the sum of the contributions at chosen points from each field would then be calculated and finally a complete diagram drawn showing the distribution of dose throughout the plane. By applying this diagram to a cross-section of the body, the dose at the tumour or at other important structures per 100 r. applied to each skin field can be found. The times of irradiation required are found from routine calibrations of the output of the x-ray apparatus. This procedure is too long and complex to be carried out in detail for each patient, but once charts are available the differences in distributions for the inevitable variations in size and shape of the patient may be estimated. The estimation of the dose at certain selected points is a routine. Until recently it was necessary to confine the axes of the beams to the same plane in order to find the distribution, but a simple apparatus has recently been devised so that the dose contours may be drawn quickly in any plane if they are known in the principal one containing the axes of the beams, so that from the standard diagrams we may now find the distributions even if the beams of radiation are inclined to each other.

Moreover, by finding the distribution of radiations in a series of parallel planes, subsequently drawing them on glass plates and placing the glass plates at the appropriate distances apart, we obtain the isodose surfaces, and so form a clear picture of the whole distribution in space even for complex arrangements of beams of radiation. These latter developments are as yet by no means routine but are quoted as recent investigations by the writer and his colleagues by which it is hoped to place radiation therapy on more certain foundations. The patient is aligned in the x-ray beams by means of an optical device which locates the beam in space and throws a spot of light on to the patient at the exit point of the axis of the beam.

It is convenient to return to the considerations of the "percentage depth dose", the clinically important measure of the penetration of the beam. This quantity is affected by a number of factors which would certainly not be allowed to enter into any physically acceptable measurements of absorption coefficient. The values of percentage depth dose depend upon the average wave-length of the radiation, but also very markedly upon the size of field of irradiation and the distance from source of radiation to skin surface. These factors have been examined in detail by many workers and tables of values of depth doses are available up to 200 kv.[39]. It may be of interest to show some of the results of measurement up to 370 kv. made by the writer, and also to compare them with those at still higher kilovoltages. Figure 3 shows the percentage depth doses along the central axis of selected beams of x rays in water. It will be seen how the values first rise below the surface and then fall rapidly with increasing depth. The values depend very greatly upon the size of area of field, and the variation of volume of irradiated tissue under the conditions normally used at from 200 to 370 kv. makes more difference to the effective penetration than even large changes in kilovoltage. The values of percentage depth dose 10 cm. below the surface are normally of the order of from 30 to 50 per cent.

It will be observed that at a depth of 4 cm. the change in the shape of the curves makes the percentage depth dose at 200 kv. and 370 kv. practically identical[40], but the difference increases rapidly at the greater depths. These tendencies have been verified at high voltages though little systematic measurement is available at potentials of the order of one million volts. Detailed studies have been made of

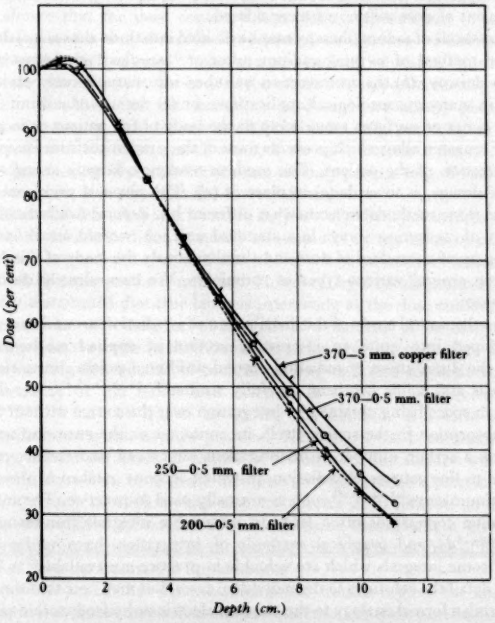


Figure 3. Focal skin distance, 122 cm. Area, 50 cm².

such questions as the relative amounts of primary and secondary radiation at various depths, and in general it may be said that at depths of the order of 15 cm. some 80 per cent of the radiation is scattered and the average wave-length of the radiation is much greater than on the surface. No complete mathematical study of the problem of depth dose has appeared, which is perhaps not surprising in view of its extreme complexity. Approximate calculations have been made[41], but there is experimental evidence that such limitations as the neglect of higher orders of scattering than the first lead to serious error.

§ 4. RADIUM THERAPY

The emphasis of this report has hitherto been upon the physical problems of high voltage x-ray therapy, but parallel developments have been taking place in γ-ray therapy. Some recent advance dependent upon the precise calculation and measurement of dose will now be considered.

The methods of radium therapy may be divided into three classes, (*a*) the so-called interstitial method of burying radium tubes or "needles" in tissues which it is hoped to destroy, (*b*) the introduction of tubes into natural body cavities, as for instance in many gynaecological applications, or (*c*) the use of radium external to the body either on surfaces moulded to fit the body of the patient or in a container of lead or tungsten alloy which prevents most of the γ radiation from escaping except in the direction of the patient. The modern tendency, largely owing to physical studies of dosage, is towards (*c*) in place of (*a*). The physical problems associated with these three methods are somewhat different but depend fundamentally on the possibility of measuring γ rays in a standard unit. A twofold attack is called for, the setting up of a standard of dose and simultaneously the study of the distribution of radiation around various types of containers. We have already dealt with the former aspect

The mathematical study of the distribution of γ radiation around various sources has developed into quite an interesting section of applied mathematics. For instance, the distribution of radiation around unfiltered points, lines, rings, radioactive shells and solids has been carefully studied[42, 43, 44, 45, 46, 47], the dosage-rate at a point in space being obtained by integration over the source without taking into account absorption in the source itself, its container or the surrounding medium. In addition a certain number of investigations have been undertaken, particularly in relation to line sources of radiation, in which account is taken of the absorption in the platinum cases[43, 44, 45, 48] such as normally used in practice. The introduction of absorption corrections often leads to intractable integrals (for instance of the form $\int e^{-\mu d/\cos\phi} d\phi$ and graphical methods of integration have to be employed. Tables of some integrals which are valuable in practice are available[19]. We cannot discuss in detail the solutions to the individual cases, but may note that the problems, while bearing a formal analogy to the classical electric or hydrodynamic calculations, are in reality quite different owing to the scalar nature of dose or dosage-rate as compared to the vectors entering the electrical problems. In some cases problems may be found with similar solutions. For instance[45], the equipotential diagrams of groups of point charges in a plane are the dose contours of infinite line sources of radiation perpendicular to the plane considered. Approximate methods of solution are often of use in clinical problems and we have drawn up tables giving the dosage-rate at various distances from point sources ranging from 0·5 mg. to 25 mg. of radium element. By breaking up the complex source into elements and by the use of the adding machine and tables it is often possible to calculate dosage-rates produced by sources having far too complex a shape for a formal solution. In

addition, the distribution of radiation around sources with uneven loading or of incomplete shells may be calculated. Recently the dose contours around solids of radioactive "fluid" have been investigated[53] by approximate methods and are useful in the theory of interstitial radiation. These mathematical studies have in fact contributed greatly to the understanding and practice of radium therapy. We may quote as an example of their utility the study of ring sources[45, 49, 50, 51]. It is easily shown that the dose contours in a plane perpendicular to and passing through the axis of a ring of radioactive material are the ovals of Cassini. From a study of these curves it appears[45] that at a perpendicular distance equal to 0·707 of its radius, the dosage-rate is the same at the centre and opposite the edge. Detailed investigation[52] shows that the deviations in dosage-rate at any place in a disc parallel to the ring at this distance are of the order of 5 per cent, so that to an accuracy sufficient for most clinical work, if we wish to treat a flat circular lesion whose surface is to receive a uniform dose, we may place a ring of radium around the periphery of a disc placed at the correct distance.

This kind of observation may be made the basis of a theory of surface applicators. In general the solution cannot be nearly so simple as the question of irradiation of healthy tissue beyond the ring and also the absolute amount of radium required to carry out the treatment are involved. However, by a combination of a ring and a uniformly distributed disc (the latter is interesting as the dose contours are ellipsoids of revolution with their foci at the ends of a diameter of the disc) it is always possible to produce a uniform distribution in a plane surface parallel to the complex source by varying the amount of radium on ring and disc correctly according to the relative size of the lesion and distance at which it is required to treat. Charts[51, 52, 53, 54] may be drawn up giving the solution in a form suitable for clinical use and are much used in practice. The subject has been much developed in recent years for all kinds of shapes of lesion both for surface work and also interstitial implantation, so that the therapist now has available precise data giving him both quantity of radium and also its correct arrangement, to deliver any previously determined minimum dose in röntgens throughout a given volume. The surface applicator work can be carried out with a higher degree of homogeneity of radiation than the interstitial implantation.

It may be objected that these mathematical studies neglect the absorption of γ radiation in the tissues themselves and are inaccurate in their practical applications on that account, but this effect for small distances and large field sizes is negligible. It is found experimentally that with wide beams of γ rays in large masses of tissues that the tissues are "transparent"[55], the scattering of the γ rays reducing the apparent absorption coefficient while the wave-length shift on scattering increases the ionization in an air-wall chamber. For narrow well-defined beams of γ rays in tissues on the other hand this is not true and work has been carried out[56] to determine the difference between the distribution of γ radiation in air and in water. It has been possible to derive apparent absorption coefficients (which are, as would be expected, approximately in accord with the theoretical values of σ_a of the Klein-Nishina formulae) which may be used to calculate more exactly the doses

in tissues at a depth[57]. The presence of bone appears to make little difference to the depth doses in the body[58], but these detailed effects have as yet been little studied. There are however available many examples of the experimental study of the dose contours around radioactive sources in water.

We have already recorded that the tendency of modern radium treatments is towards the use of external sources rather than interstitial radium, and one of the important consequences has been the development of units for such use. Originally such units (sometimes unfortunately termed "bombs", or in America "packs") consisted simply of a hollow lead block in the centre of which the radium was placed, an aperture in the lead defining the beam of γ rays. In general a wall thickness of about 5 cm. of lead is required to reduce the radiation from 1 g. of radium to reasonable limits in the unwanted directions. Many units of this simple type have been used for some years, as for instance the famous unit at the Radium Institute, Paris[59], which contained 4 g. of radium in a flat lead box with a defining aperture. The lead containers may weigh several hundredweights and are supported on mechanical devices so that the unit may be raised and lowered, moved laterally as well as rotated on an axis to allow of accurate positioning with respect to the patient. Another well-known radium pack of more elaborate design is that due to Failla[60] in which the radium is mounted on a lead cylinder which rotates on the centre of the main lead block. This cylinder may be moved so as to cut off the beam of γ rays when required or to bring the radium opposite an aperture of smaller dimensions. Many other designs are extant but we will limit ourselves to the description of an apparatus recently developed[61] at the Radium Beam Therapy Research and which represents probably the most convenient and safest form of unit now available.

The radium unit is constructed partly of lead, but the nozzles which are applied to the patient, and the head of the unit are composed of an alloy of tungsten, copper and nickel[62] with a density of about 16·5, that is about 50 per cent greater than that of lead. Since for heavily filtered γ rays the density is the most important factor in determining absorption coefficients, the protective thickness of alloy are less than those in lead. This is an important matter in practice as one of the main difficulties is to make an apparatus which may be fitted into or against the complex shapes of the human head and neck (the region chiefly treated by this method) while at the same time retaining sufficient protection of the remainder of the patient's body and giving a well-defined beam of γ rays.

The unit itself which has admirable mechanical movements is connected by a flexible metal tube with a well protected safe containing lead blocks affording some 15 in. protection in almost all directions. The radium is mounted on a bobbin of light alloy which is blown from safe to unit by means of a commercial vacuum cleaner when the treatment is to commence. By a suitable valve it is easily arranged for the radium to be blown back again to the safe at the end of the treatment, while a system of light signals shows the position of the radium at any time. An electric clock is also incorporated which times the treatment and in later models closes ontacts which returns the radium automatically when desired. The apparatus is

controlled from a distant cubicle and the protection of the staff is excellent as all manipulations are carried out with the radium in the safe.

It is impossible to discuss in detail the problems connected with the distribution of radiation in the tissues from such units, but they raise some points of interest in comparing the use of radium and X rays from the physical point of view[63].

By the methods outlined above, that is the use of an air-wall chamber in a water phantom, it is possible to determine the distribution of radiation in water and express it by means of dose contours as for X rays. An example is given in figure 4, for one field of a new type of unit in use at the Royal Cancer Hospital.

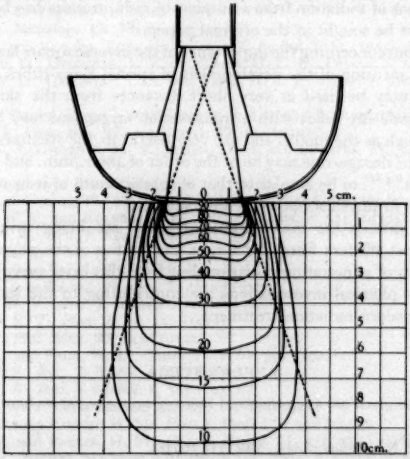


Figure 4. One-gramme radium unit. Surface dosage rate, 180 r./hr.; diameter of field, 3 cm.; focal distance, 6 cm.

It is found that the percentage depth doses even with a 5 g. unit are small compared to the values obtainable with high voltage X rays. For instance, for a 400 kv. apparatus the dose 10 cm. below the surface may rise, for reasonable field sizes, to approximately 50 per cent of that on the surface, whereas for a 5 g. unit the value would probably be about 15 per cent. The reason lies of course in the very different focal-skin distances, which reflect the absolute amount of radiation available in the two cases. Now that we may measure both radiations in röntgens the difference is very apparent. With 5 cm. focal-skin distance a 1 g. unit will give 200 r./hr. or about 3·3 r./min. It is not practicable to work much below this figure as the dose per session is often of the order of 300 r. and an hour and a half is about the maximum length of time for which a patient may be kept still with reasonable comfort. On the other hand with a 200 kv. X-ray apparatus at 1 m. focal distance the dosage-

rate from a constant potential plant may easily reach 30 r./min. It is therefore necessary to bring even large radium sources (say 5 g.) fairly near the surface (about 8 cm.) and the inverse square law then makes large percentage depth doses impossible however penetrating the electromagnetic radiation.

We have therefore the at first sight curious result that even the normal 200 kv. plant yields radiation of effectively greater penetration under clinical conditions than the radium sources. The same applies *a fortiori* to X-ray apparatus at still higher voltages of the order of a million when the focal distance may if desired be still further increased and depth doses of 60 per cent at 10 cm. depth are attainable. The distributions of radiation from a number of radium units has been determined but details must be sought in the original papers[57].

This last point concerning the importance of the inverse square law leads naturally to a very short mention of the development of special X-ray tubes for use at 50 or 60 kv., which may be used at very short distances from the skin[64]. They are essentially cathode ray tubes with a transmission target and may be inserted into body cavities such as the mouth and are very useful in the treatment of superficial cancer. Here the dosage-rate may be of the order of 400 r./min. and the distribution has been shown[65,66] to be similar to that of a teleradium unit in spite of the very great difference in physical nature of the radiation.

We cannot discuss the many problems which are being investigated, as for instance the most efficient filters for X rays, the studies of the physical basis of the biological effects of penetrating radiation, but from this brief survey it will perhaps be realized that physical investigations are contributing to the success, accuracy, and safety of modern radiation treatment.

REFERENCES

The references given are a selection and in many cases the papers have extensive bibliographies.

(1) Mackenzie, K. *Nature, Lond.*, **142**, 116 (1938).
(2) Murdoch, J., Simon, S. and Stahel, E. *Acta radiol., Stockh.*, **11**, 350 (1930).
(3) Kroenig, B. and Friedrich, W. *Physikalische und biologische Grundlagen der Strahlentherapie* (1918).
(4) Behnken, H. *Z. tech. Phys.* **5**, 3 (1924).
(5) Behnken, H. *Strahlentherapie*, **29**, 192 (1928).
(6) Duane, W. *Amer. J. Roentgenol.* **19**, 461 (1928).
(7) Taylor, L. S. *Radiology*, **16**, 1 (1931).
(8) Kaye, G. W. C. and Binks, W. *Brit. J. Radiol.* **2**, 553 (1929).
(9) Failla, G. *Amer. J. Roentgenol.* **21**, 47 (1929).
(10) Mayneord, W. V. and Roberts, J. E. *Brit. J. Radiol.* **7**, 158 (1934).
(11) Taylor, L. S. *Bur. Stand. J. Res.* **8**, 9 (1932).
(12) Fricke, H. and Glasser, O. *Fortschr. Röntgenstr.* **33**, 239 (1925).
(13) Gray, L. H. *Proc. Roy. Soc.* A, **124**, 647 (1929).
(14) Gray, L. H. *Brit. J. Radiol.* **10**, 600, 721 (1937).
(15) Glasser, O. and Mautz, F. R. *Radiology*, **15**, 93 (1930).
(16) Mayneord, W. V. *Brit. J. Radiol.* **4**, 693 (1931).
(17) Mayneord, W. V. *Brit. J. Radiol.* **6**, 598 (1933).
(18) Mayneord, W. V. and Roberts, J. E. *Brit. J. Radiol.* **10**, 365 (1937).

(19) Sievert, R. *Acta Radiol., Stockh.*, Supplementum, **14** (1932).
(20) Eve, A. S. *Phil. Mag.* **22**, 551 (1911).
(21) Eve, A. S. *Phil. Mag.* **27**, 394 (1914).
(22) International Recommendations for Radiological Units. *Brit. J. Radiol.* **10**, 438 (1937).
(23) Failla, G. and Henshaw, P. S. *Radiology*, **17**, 1 (1931).
(24) Kaye, G. W. C. and Binks, W. *Proc. Roy. Soc.* A, **161**, 564 (1937).
(25) Lauritsen, C. C., *Brit. J. Radiol.* **11**, 471 (1938).
(26) Mayneord, W. V. *Acta Union Int. Cancer*, **11**, 271 (1937).
(27) Mayneord, W. V. *Proc. Roy. Soc.* A, **130**, 63 (1930).
(28) Failla, G. and Marinelli, L. D. *Amer. J. Roentgenol* **38**, 312 (1937).
(29) Lauritsen, C. C. *et al. Phys. Rev.* **32**, 850 (1928); **36**, 988 (1930).
(30) Allibone, T. E. *Engineer* (16 April 1937).
(31) Bouwers, A. *Radiology*, **22**, 163 (1934).
(32) Bouwers, A. and van der Tuuk, J. H. *Brit. J. Radiol.* **7**, 21 (1934); **9**, 431 (1936).
(33) Coolidge, W. D., Dempster, L. E. and Taris, H. E. Jr. *Physics*, **1**, 230 (1931).
(34) Coolidge, W. D., Dempster, L. E. and Taris, H. E. Jr. *Amer. J. Roentgenal.* **27**, 405 (1932).
(35) Dresser, R., Trump, J. and Van de Graaff, R. J. *Amer. J. Roentgenal.* **38**, 758 (1937).
(36) Sloan, D. H. *Phys. Rev.* **43**, 213 (1933); **47**, 62 (1935).
(37) Stone, R. S., Livingston, M. S., Sloan, D. H. and Chaffee, M. A. *Radiology*, **24**, 153 (1935).
(38) *Reports on Progress in Physics*, **3** (1936).
(39) Grebe, L. and Nitzge, K. *Strahlentherapie*, Sonderband, **14** (1930).
(40) Mayneord, W. V. and Roberts, J. E. *Brit. J. Radiol.* **6**, 321 (1933).
(41) Payne Scott, R. *Brit. J. Radiol.* **10**, 850 (1937).
(42) Hess, V. *Phys. Rev.* **19**, 73 (1922).
(43) Sievert, R. *Acta Radiol., Stockh.*, **1**, 89 (1921).
(44) Sievert, R. *Acta Radiol., Stockh.*, **11**, 249 (1930).
(45) Mayneord, W. V. *Brit. J. Radiol.* **5**, 677 (1932).
(46) Schweidler, E. V. *Phys. Z.* **13**, 453 (1912).
(47) King, L. V. *Phil. Mag.* **23**, 242 (1912).
(48) Owen, E. A. and Fage, W. E. *Chem. Tr. J.* **69**, 791 (1921).
(49) Souttar, H. S. *Brit. J. Radiol.* **4**, 681 (1931).
(50) Souttar, H. S. *Brit. J. Radiol.* **8**, 385 (1935).
(51) Mayneord, W. V. *Acta Radiol., Stockh.*, **14**, 95 (1932).
(52) Paterson, R. and Parker, H. M. *Brit. J. Radiol.* **7**, 592 (1934).
(53) Paterson, R. and Parker, H. M. *Brit. J. Radiol.* **11**, 252, 313 (1938).
(54) Paterson, R., Parker, H. M. and Spiers, F. W. *Brit. J. Radiol.* **9**, 487 (1936).
(55) Bruzeau, M. *Ann. Phys., Lpz.*, **5** (1929).
(56) Roberts, J. E. and Honeyburne, Joan M. *Brit. J. Radiol.* **10**, 515 (1937).
(57) Mayneord, W. V. and Honeyburne, Joan M. *Brit. J. Radiol.* **11**, 741 (1938).
(58) Wilson, C. W. *Brit. J. Radiol.* **9**, 38 (1936).
(59) Ferroux, R. and Bruzeau, M. *Arch. Inst. Radiol.* **1**, Fasc. 2, 324.
(60) Failla, G. *Amer. J. Roentgenal.* **20**, 128 (1928).
(61) Grimmett, L. G. *Brit. J. Radiol.* **10**, 105 (1937).
(62) McLennan, J. C. and Smithells, C. J. *J. Sci. Instrum.* **12**, 159 (1935).
(63) Mayneord, W. V. *J. Sci. Instrum.* **12**, 245 (1936).
(64) Chaoul, H. *Strahlentherapie*, **48**, 31 (1933).
(65) Ernst, H., Frik, K. and Ott, P. *Strahlentherapie*, **52**, 389 (1935).
(66) Mayneord, W. V. *Brit. J. Radiol.* **9**, 215 (1936).

ABSOLUTE ELECTRICAL MEASUREMENTS

By L. HARTSHORN, D.Sc.

All measurement is conventional, and it would be possible to devise a perfectly serviceable system of measurement, according to which a man would be larger than the sun.

BERTRAND RUSSELL, *An Outline of Philosophy*

§ 1. INTRODUCTION

In recent years various Conferences and Committees of the International Bureau of Weights and Measures have considered the position of electrical units and standards, and it has been decided that, beginning in 1940, the present International Units shall be abandoned, and units based on "the absolute system" adopted for all purposes. At a meeting held in Paris in June 1937, the following provisional values[1] for the new units were agreed upon:

1 international ohm = 1·000 48 Ω. (absolute),

1 international volt = 1·000 36 v. (absolute).

Changes of this order will be of importance to many experimental physicists, and it is therefore proposed to review the work on which these pronouncements are based, with the object of making clear the basis of the new system, and the significance of the changes.

§ 2. MEANING OF "ABSOLUTE"

The term "absolute units" is sometimes used in a way which suggests that such units are held to be the natural ultimate basis of all physical measurements, and that the use of any other system of units is merely due to ignorance. Such a view will not bear close examination. It can only arise from a confusion of the ideas associated with the word "absolute" in its more general sense with those properly associated with this special use of the word. The limited sense in which the word is used was clearly defined by the British Association's Committee on Electrical Measurements in their *Report*[2] for the year 1863, which states "The word 'absolute' in the present sense is used as opposed to the word 'relative' and by no means implies that the measurement is accurately made, or that the unit employed is of perfect construction; in other words, it does not mean that the measurements or units are absolutely correct, but only that the measurement, instead of being a simple comparison with an arbitrary quantity of the same kind as that measured, is made by reference to certain fundamental units of another kind treated as postulates." The B.A. Committee was not responsible for the introduction of the word "absolute". It is evident from this *Report* that the word was in fairly general use before the Committee was formed, and that they realized that the term was in some respects an unfortunate one. They

state "It appears probable that the name of 'derived units' would more readily convey the required idea than the word 'absolute', or the name 'mechanical units' might have been adopted; but when a word has once been generally accepted, it is undesirable to introduce a new word to express the same idea"[3]. There can be little doubt that the use of the Committee's term "derived units" would have avoided misunderstanding as to the object of the system. On this point the Committee is quite definite. "The object or use of the absolute system of units may be expressed by saying that it avoids useless coefficients in passing from one kind of measurement to another."[3] As long as this moderate claim is borne in mind the nature of absolute units will not be misunderstood.

§ 3. THE ABSOLUTE SYSTEM OF UNITS

From the above definition of the term "absolute" given by the British Association Committee, it follows that many absolute systems of units could be constructed. An absolute measurement is simply the deduction of the magnitude of one quantity from observations of the magnitudes of other quantities of a different nature. Such a deduction is obviously only possible when these quantities are related in a way which may be expressed by a quantitative law which is accurately known. It is necessary to have a series of such laws connecting all the quantities to be measured either directly or indirectly with those fundamental quantities whose units are to be "treated as postulates". Obviously the system of units obtained in this way will depend not only on the fundamental units selected, but also on the particular quantitative laws chosen. It is well known that the B.A. Committee chose the units of length, mass, and time as their fundamental units, and Ampere's Law, Faraday's Law of Induction, Ohm's Law, and the law of the equivalence of electrical and mechanical energy, or various general equations which are regarded as equivalent to these laws as the relations governing their "absolute system of units", and when the International Committee of Weights and Measures decided to adopt a system, based on "the absolute system", it was generally understood that this was the system referred to. Nevertheless, it should be understood that this particular choice of laws and fundamental magnitudes was quite arbitrary, and that "the absolute system" of units is no less arbitrary than any other system, though it is undoubtedly more convenient.

The units of the absolute system are usually held to be defined by the general equations expressing the above-mentioned laws, and these are expressed in various well-known forms which are theoretically all equivalent. When verbal definitions are considered desirable, they are usually given by describing some simple system to which the law is applicable. Many such definitions are possible but they are all considered to define the same unit, the general equation, and the enumeration of the fundamental units (e.g. c.g.s.) constituting the real definition. Examples of such verbal definitions are to be found in Appendix C of the B.A. *Report* already quoted. Thus we find "The unit of current is that current of which the unit of length placed along the circumference of a circle of unit radius, produces a unit of magnetic force

at the centre"[4], or if it is considered desirable to avoid the use of magnetic quantities "the unit electric current conducted round two circuits of unit area in vertical planes at right angles to each other, one circuit being at a great distance D vertically above the other, will cause a couple to act between the circuits of a magnitude $1/D^2$".[5] In this way the unit of current is defined in terms of the units of mass, length, and time. The unit of potential difference is then defined in terms of the unit of current and the unit of power, the unit of quantity in terms of the units of current and time, the unit of resistance in terms of the units of current and p.-d., the unit of inductance in terms of the units of current, p.-d. and time, and so on. The definitions are too familiar to require quotation. In the words of the B.A. *Report* "all the units form part of a coherent system" which is based on the fundamental units of mass, length and time. It follows that it is possible to deduce the magnitude of any of these quantities from experiments involving measurements of mass, length and time alone, and such a process constitutes a measurement in "the absolute system" of units.

We have seen that the alternative to the absolute system is the *relative system* in which each quantity is simple compared with another quantity of the same kind which is chosen as standard. Such measurements are usually far more accurate than absolute measurements, but the necessity of maintaining a separate standard for each of the quantities to be measured constitutes a serious disadvantage, for in practice it is not possible to construct standards that do not vary appreciably with time, and it therefore becomes very difficult to relate measurements made at different times. One of the main objects in adopting the absolute system of units was to overcome this difficulty. Absolute measurements have always been capable of less precision than relative ones, but since the existing standards of length, mass and time are more permanent than any electrical standard hitherto constructed, absolute measurements made at widely different times are likely to be more consistent than relative measurements.

§4. THE INTERNATIONAL POSITION: NOMENCLATURE

International agreement as to the units to be employed for electrical measurements has been obtained at various conferences, the last of which was held in London in 1908. This Conference adopted fundamental units based on "the absolute system" for all electrical measurements in the following terms:[6]

"The Conference agrees that as heretofore, the magnitude of the fundamental electric units shall be determined on the electromagnetic system of measurement, with reference to the centimetre as the unit of length, the gramme as the unit of mass, and the second as the unit of time. These fundamental units are:

(1) The ohm, the unit of electrical resistance, which has the value of 1,000,000,000 in terms of the centimetre and second;

(2) the ampere, the unit of electric current, which has the value one-tenth (0·1) in terms of the centimetre, the gramme and the second;

(3) the volt, the unit of electromotive force, which has the value 100,000,000 in terms of the centimetre, the gramme, and the second;

(4) the watt, the unit of power, which has the value 10,000,000 in terms of the centimetre, the gramme and the second."

From this resolution it appears that the ohm, ampere, volt, and watt, here clearly defined as absolute units, have been internationally recognized as the fundamental units since 1908. The actual wording of the definition is of some interest in connexion with nomenclature. It is frequently stated that the ohm is equal to 10^9 cm./sec., but such statements are apt to be very misleading, and were carefully avoided in the resolution. The absolute unit of resistance is itself a resistance and not a velocity. The value of a resistance in absolute units may be deduced from the measurement of a velocity on the assumption that the laws on which the absolute system is based are valid, but the definition of an absolute measurement stresses the fact that the resistance and the velocity are quantities of a different nature, and it is therefore misleading to give unit resistance and unit velocity the same name. In the same way it is very doubtful whether the common practice of calling the unit of inductance the centimetre can be justified. The abbreviation c.g.s.m.u. which indicates the fundamental units (c.g.s.) and the fundamental equations (*m*) is satisfactory for all the units of the system which have no special name.

Although the London Conference of 1908 adopted the absolute units as the fundamental basis of electrical measurement, it was not considered practical to rely entirely on absolute measurements for standardizing purposes. Relative measurements are so much more convenient and more accurate that there could be no question of dispensing with them. It therefore was necessary to construct standards and calibrate them in terms of the units adopted. Experience shows that the most convenient standards for most purposes are resistors and standard cells: others are very easily derived from them. But in 1908 it was not considered possible to construct standard resistors and standard cells which would be sufficiently stable to be regarded as primary standards. The obvious thing to do was to calibrate such standards periodically by the most accurate absolute measurements available, but the Conference as a whole had not sufficient confidence in the convenience and accuracy of absolute methods to accept this plan. It was considered very desirable that every national laboratory should be able to reproduce the primary standards at any time and therefore standards were adopted which were believed to be superior to any others then available, as regards ease and accuracy of reproduction. The resolutions of the Conference which followed the first resolution already quoted were as follows:

"(ii) As a system of units representing the above, and sufficiently near to them to be adopted for the purpose of electrical measurement, and as a basis for legislation, the Conference recommends the adoption of the international ohm, the international ampere, and the international volt defined according to the following definitions:

(iii) The ohm is the first primary unit.

(iv) The international ohm is defined as the resistance of a specified column of mercury.

(v) The international ohm is the resistance offered to an unvarying electric current by a column of mercury at the temperature of melting ice 14·4521 g. in mass of a constant cross sectional area and of a length of 106·300 cm.

(vi) The ampere is the second primary unit.

(vii) The international ampere is the unvarying electric current which, when passed through a solution of nitrate of silver in water, deposits silver at the rate of 0·00111800 g./sec."

The international volt, coulomb, watt, joule, henry, and farad were derived from the above units by definitions exactly similar to those by which the corresponding units in the absolute system are derived. These international units have formed the basis of all ordinary electrical measurements since that time. It should be noticed that they are defined to the nearest 1 part in 10^5: it follows that values cannot be expressed in international units with an accuracy greater than this. As mentioned in the introductory paragraph, it is now known that these units differ from the corresponding absolute units by a few parts in 10^4.

§ 5. LEGAL STANDARDS

The object of the Conference in devising the above definitions was to define accurately reproducible units. The standards described in the definitions have, strictly speaking, never been constructed. It would clearly be a physical impossibility to get a column of mercury of just those dimensions, although one can, by a series of measurements on columns of nearly those dimensions, deduce with sufficient accuracy the resistance of the specified column, but a standard which does not actually exist is not likely to commend itself to the legal faculty. The legal standard of resistance in this country is "the resistance between the copper terminals of the instrument marked 'Board of Trade Ohm Standard Verified 1894 and 1909' to the passage of an unvarying electrical current when the coil of insulated wire, forming part of the aforesaid instrument and connected to the aforesaid terminals, is in all parts at a temperature of 16·4°c.", and this resistance is for the purposes of the Weights and Measures Act, 1889, "denominated one Ohm." The limits of accuracy attainable in the use of the said denomination are stated to be within one hundredth part of 1 per cent, and within these limits this standard agrees with the international ohm. Similarly the legal standard of current is the current passing through the coils of a particular ampere balance and balancing a particular iridioplatinum weight, while the legal standard of voltage is one hundredth part of the voltage which, when applied to a particular electrostatic voltmeter, causes it to pass from one specified condition to another specified condition, i.e., causes it to give full scale deflection. These two standards also agree with the international units within the limits of accuracy prescribed for them (0·1 per cent).

§ 6. WORKING STANDARDS

It is now possible to make relative measurements of resistance—to compare the resistances of standard manganin coils—with an accuracy better than 1 part per million and it will be found that recent reports of the International Committee of Weights and Measures give values of resistance to 0·1 part per million[7]. Such values cannot be expressed in international units, for these units have only been

defined to 10 parts per million. It is clearly desirable to have a recognized unit of some kind for such measurements, and in practice each standardizing laboratory adopts a "working unit". Coils calibrated in these units are compared at the International Bureau of Weights and Measures, and as a result of the comparisons, a "mean international unit" is determined, and values of all the coils are expressed in terms of this unit. It is understood that the units determined from time to time in this way are probably not constant, and that they have not been officially defined or adopted for general use. Nevertheless they give a useful guide to the behaviour of various standard coils, within limits finer than those obtainable in the use of the international units. Similarly standard cells are compared with an accuracy of 1 part per million and a mean unit of electromotive force is deduced from comparisons of standard cells calibrated in terms of the working units of the various national laboratories[8]. These working units are arbitrary units adopted by the national laboratories on the basis of their own determinations of the international units, and the behaviour of their own standards as judged by relative measurements. For all ordinary purposes they may be regarded as identical with the international units.

§ 7. THE EXPERIMENTAL BASIS OF THE ABSOLUTE SYSTEM

In the foregoing paragraphs the absolute system of units has been outlined in the terms employed by the founders of the system and generally accepted by physicists and others who use it. In a valuable recent paper Dr Norman Campbell[9] has pointed out that this traditional exposition of the subject conceals difficulties and complexities that cannot be ignored if we wish to know with precision the real significance of our measurements. We have seen that the absolute system is based on certain generalizations which are usually represented by the fundamental equations of electromagnetism. But these beautifully simple generalizations are theories rather than experimental facts, and it is desirable that the essentially practical business of measurement should be based on facts rather than theories. Of course the theories and general equations have themselves been derived from, or rather suggested by, experimental facts, but the derivation has included a good deal of purely imaginative work, and most students of physics have some difficulty in convincing themselves that the fundamental equations have actually been proved experimentally. It is obvious that a great body of experimental facts has been established with great precision and it is vaguely felt that these facts justify both our systems of measurement and our fundamental working equations. Dr Norman Campbell has discussed the nature of the relationship that must exist between the proposed absolute system of units and those experimental laws which have been established with the precision that is essential for the basis of such a system, and has suggested a possible scheme whereby the system might be formulated in a systematic and logical manner on a basis of established facts. Dr Campbell's argument is naturally less simple than one which assumes the simple theoretical generalizations, but in view of its importance to all physicists who are concerned with the real basis of their measurements, it will be reproduced in the following paragraphs in sufficient detail to indicate its character.

It is first necessary to consider the way in which an experimental law may be established and used for the purposes of measurement. We must have an apparatus which is characterized by two measurable properties x and y, and an identifiable state P (e.g. of equilibrium) which is assumed for a given value of x when and only when y has some particular value. We find a number of identifiable systems S possessing some distinguishing property. We include these systems in turn in the apparatus and find the pairs of values x and y for which the system assumes the state of reference P. We then find that there is a relation between the pairs of values x, y and the distinguishing property of the corresponding systems which may be expressed by the algebraic equation

$$f(xy) = C, \qquad \ldots\ldots(1)$$

where C varies with the system S inserted, C_r being characteristic of the member S_r. Then C is a property of the systems S which is measurable in virtue of this relationship.

The value of C_r depends on the units used in measuring x and y, a fact which may be expressed by writing the equation in the form

$$f(xy) = C'Z \qquad \ldots\ldots(2)$$

and adopting the convention that when the units of x and y are changed, only C' is changed. Then Z_r is a characteristic of the system S_r only. This convention alone does not settle the value of C', for many values could be chosen which would satisfy it. The ambiguity may be resolved in two ways. We may adopt what may be called the absolute principle, and choose arbitrarily one pair of units of xy, and assert that C' has some simple definite value, usually 1, for these units. C' is thus suppressed and the object of the absolute system, the avoidance of "useless coefficients", is achieved. Alternatively, we may choose arbitrarily one of the systems S, say S_1, and assign to it some value Z_1, say 1; then C' is determined by the experimental facts.

No complete system of measurement can be established by such processes alone, for the process implies that some properties x and y are already measurable. Somewhere in the background must lie another process, depending on laws that are not numerical. These are always laws of physical addition, which state that certain combinations of two systems S are equivalent in some special sense to a third member. The establishment of these laws does not involve measurement; but they form the basis of all measurement.

Starting with this conception of the processes of measurement, the following scheme is given as a possible basis, independent of electrical theories, for the absolute system of units which it is proposed to adopt. The actual basis will clearly depend on the kind of apparatus which proves to be usable with the greatest precision, for the relation expressed by equations such as (1) imply not only apparatus of a particular nature but also a definite state of reference P. It is scarcely possible to give an adequate summary of the whole scheme, but the initial stages are specially interesting on account of their contrast with the traditional treatment of the problem.

For brevity the mechanical units are taken for granted; it is assumed that we can measure length, time, mass, force, power, angle, area. As three initial arbitrary assumptions it is asserted that the units of length, mass and time are the metre, kilogramme, and second. (The centimetre, gramme, and second would of course do equally well, except that coefficients of high powers of 10 would appear in many equations if the ohm, ampere, volt, etc., are to be the units.)

The systems first dealt with are pairs of linear conductors surrounded by non-magnetic material. (Note that it would be possible to discriminate between conductors and non-conductors, magnetic and non-magnetic material, without the necessity of measurement: detecting instruments, not measuring instruments would be used.) A quantity M is first defined by the familiar relation

$$M = 10^{-7} \iint \frac{\cos \epsilon \, ds \, . \, ds'}{r} . \qquad \ldots\ldots(3)$$

This quantity can clearly be determined by measurements of lengths and angles. We make the assumption that M is a physical property characteristic of the systems, and the unit of M is determined by this assumption. It is to be noted that $dM/d\theta$, where θ represents rotation of one conductor relative to the other, and dM/dz where z represents relative translation, are also measurable for a determinate value of θ or z.

The first experimental law concerns resistors, and is derived from observations with the Lorenz apparatus. The measurable properties x and y are $dM/d\theta$ and w the relative angular velocity of the two conductors to which M appertains. The state of reference P is the condition of no galvanometer deflection when a source of current is connected in series with one of a number of resistors (the systems S) and one of the linear conductors, the galvanometer being connected in series with the other conductor, and this combination joining the terminals of the resistor. By measurements of w the relative angular velocity of the two linear conductors and $dM/d\theta$ we establish the law

$$w \, . \, dM/d\theta = C. \qquad \ldots\ldots(4)$$

Then assuming that C is to be written $1 \times R$, we derive R (ohms) as a measurable property of the resistors. It should be emphasized that no knowledge of the nature and properties of electric currents is required for this process. Thus the ohm may be regarded as the first fundamental electric unit, and it is perhaps significant that the London Conference of 1908 passed a resolution to this effect (see p. 305).

The above law has been established for only a very limited range of R. Practical resistance measurement depends mainly on the additive properties of R, which are established by the following process. We can find three resistors S_1, S_2, S_3 such that if S_1 and S_2 are connected in series they are equivalent to S_3 in the sense that the substitution of one for the other produces no change in a suitable detecting apparatus. In virtue of this fact we find we can assign numerals S to the resistors consistently, so that for such resistors $S_1 + S_2 = S_3$. If we assign the numeral 1 arbitrarily to one resistor, then the corresponding numerals for the others serve to measure a property R' of the resistors, quite independently of equation (4) or any

other numerical law. We now find that in the range for which (4) has been established R' is proportional to R, and if we use for R' the unit which has been established for R, the two quantities R' and R become identical. By similar experiments we can discover the additivity of conductance G and identify G with $1/R$. These are the laws on which resistance measurement is based, the previous law (4) is used merely to establish the unit. It should be noticed that the process does not depend on Ohm's Law.

The next experimental law is that derived from experiments with current-balances. The systems S are in customary terms conductors carrying currents, but in order that they may be identified and distinguished they may include silver voltameters. The mass of silver deposited in a given time can then be measured and will serve for identification purposes. The measurable properties xy of the balance are F, the force applied in a direction z between two linear conductors forming part of the balance, and dM/dz for these conductors. The state P is that of equilibrium when the voltameter circuit and the two conductors are all connected in series. The law established by the experiments is

$$F = C.dM/dz. \qquad \text{.....(5)}$$

We then assume that C may be written $1 \times i^2$ and i (amperes) becomes a measurable property of the systems. The choice of the assumption $C = i^2$ rather than $C = i'$ could be justified by showing that the i of the first assumption is additive in branched circuits, whereas the i' of the second assumption is not.

The next experimental law is that derived from measurements with the potentiometer. By measurements on voltaic cells the law

$$Ri = C \qquad \text{.....(6)}$$

is established, and on the absolute principle we write $C = 1 \times E$, and E (volts) is a measurable property of the cells (potential difference).

The laws derived from alternating current bridges then serve for the measurement of inductance and capacitance. They may be written in the familiar notation

$$jwL = R, \qquad \text{.....(7)}$$

$$-\frac{j}{wC} = R. \qquad \text{.....(8)}$$

The whole series of electrical and magnetic units is derived by the systematic application of similar processes. The arguments based on theoretical generalizations are of course very much simpler and this justifies their use for everyday laboratory work, but if a new attempt is to be made to define the system with the maximum possible precision, a strong case can be made for the avoidance of theories which in the light of later development may prove to be no more than approximations, and the adoption of a basis of experimental fact with certain well-defined assumptions. The examples quoted show that in order to define the five units for R, i, E, C, M (and L), five arbitrary assumptions are necessary, and each additional unit requires another assumption, facts which are often overlooked.

Dr Campbell's treatment of the subject is undoubtedly much more complicated than the usual one based on the simple fundamental equations of electromagnetic

theory, and we shall therefore continue to discuss the details of our measurements in terms of theory, but a survey of the activities of the standardizing laboratories will show that these depend almost entirely on observations made with calculated inductance standards, ampere balances, resistance build-ups, potential dividers, and alternating current bridges, and an analysis of this kind of the exact relation between the observations made and the system of units finally arrived at is very desirable at a time like the present when a new system of units is to be adopted.

Recent developments in the practice of the measurements will now be discussed.

§8. PRIMARY STANDARDS OF INDUCTANCE

Whatever system may be employed for the definition of the absolute electromagnetic units, it is an undoubted fact that the starting point of all the precision determinations of the units so far carried out is the construction of a standard of inductance, either self or mutual. The problem is to construct a coil or coils of wire of such a form that their linear dimensions may be determined with the highest possible accuracy, and the inductance calculated from them by an equation of the type (1). All existing standards of high precision include as their most essential part a single-layer solenoid wound on a cylindrical former. The problem of constructing and measuring such a coil by reference to standards of length would at first sight appear to be a relatively simple one, but in some of the most recent investigations it is estimated that the errors involved in such measurements and calculations are greater than those involved in the purely electrical measurements in which such standards are subsequently used. The diameter of a cylinder of uniform diameter can easily be measured with the accuracy required, better than 1 part per million, but unfortunately it is a very difficult matter to construct a helix of uniform diameter and uniform pitch, on a former of non-magnetic and non-conducting material. Irregular variations of both diameter and pitch are appreciable in all methods of construction so far developed, and it follows that not only is it necessary to make measurements at a very great number of points along the helix, but the evaluation of the integral (1) becomes more uncertain the more irregular the variations in the linear dimensions. The integral can fairly easily be evaluated for a uniform helix by various well-known formulae, and in practice these formulae are evaluated for a hypothetical helix of a diameter and pitch equal to the mean of all the measured values of the actual helix. This calculation can be made with the accuracy required, but the differences between the value so obtained and the value for the actual helix with its irregularities of construction can only be estimated approximately by treating the various turns of the helix as equivalent to circles of a diameter and axial position equal to the mean measured value for the corresponding portion of the helix. It will be obvious that the accuracy obtained by such a procedure will be better the more nearly uniform is the helix, and it follows that a great amount of research has been done recently, particularly at the National Bureau of Standards, Washington, on the construction of coils of this type. The earlier standards, such as those of the National Physical Laboratory, constructed in 1907,

and of the Physikalisch-technische Reichsanstalt, Berlin, were wound on cylinders tuned from statuary marble, bare copper wire being wound directly on to the marble cylinder, and the turns in the N.P.L. instrument being located by helical grooves cut into the surface of the cylinder. The wire was wound under tension, so that the temperature coefficient of expansion of the coil is equal to that of the marble, which is small (about 4×10^{-6} per 1°C.) Marble was used as being rigid but machinable, and a sufficiently good insulator. However, its rigidity is far from perfect. Lord Rayleigh[10] has recently shown that although it does not actually flow under moderate pressures, its flexure under load is not entirely elastic, and if the material is heated (as has sometimes been done to improve the insulating properties) the structure is liable to become loosened owing to unequal expansion in different directions of the grains of calcite, and both the mechanical properties and the linear dimensions are changed. The insulating properties of marble are also inferior to those of many other materials, and the American workers[11] have therefore turned to porcelain, glass of the Pyrex type, and fused quartz. There can be little doubt that fused quartz is the best of these materials, in dielectric and mechanical properties, as well as on account of its very low temperature-coefficient of expansion. Only in recent years has it been possible to produce it in large sizes, and it is difficult to work, but inductance standards wound on cylinders of fused quartz have now been successfully manufactured in both America[11] and France[12], and it is probable that this material will in time supersede the others for such work.

Fused quartz cannot be turned and the cylinder must therefore be ground, but by the use of the technique of the optical workshop, Dr Moon[11] of the National Bureau of Standards has succeeded in producing a cylinder of diameter 28 cm. of which the variations in diameter were less than a micron or 4 parts per million. The process is essentially simple but requires great care and patience. The roughly ground cylinder is mounted so that it can be turned slowly about a horizontal axis. It is then covered with a mixture of carborundum and water and lapped by lightly pressing a hand-controlled saddle against the cylinder. The actual grinding is done by two grooved aluminium bars mounted in a light frame forming the above-mentioned saddle, which rests on the surface of the cylinder, the bars being parallel to the axis and separated by about 90° of arc. The bars are about two-thirds of the length of the cylinder. The lapping is performed by sliding the saddle slowly and uniformly from end to end of the cylinder, but with a stroke that carries about half the length of the bars beyond the end of the cylinder. After a certain number of strokes the saddle is turned end for end, and the grinding continued for the same number of strokes. At frequent intervals the cylinder is cleaned and the diameter measured. If the centre is found to be grinding more rapidly than the ends, the length of stroke is increased and vice versa. Local irregularities were removed by the use of a shorter lap of similar construction. The long and short laps were used in turn with finer and finer grades of abrasive until the desired result was obtained. In his case the wire was wound on to the smooth face of the cylinder, the under-side of the wire being coated with a trace of shellac-varnish to act as an adhesive. Uniformity of diameter of a high order was achieved in this way, but the winding

was not so satisfactory as regards uniformity of pitch. Still more recently Dr Moon[13] has used a similar technique for the cutting of a spiral groove in the surface of a glass cylinder. The results are reported as encouraging and there is good reason to expect that a coil formed by winding wire into such a groove would be superior to any previously constructed.

Even if the cylinder were perfectly uniform, irregularities in the coil might arise from irregularities in the diameter of the wire used. Wire as ordinarily supplied is of elliptical rather than circular cross-section, and irregularities of diameter of the helix often arise from the changes in the orientation of the elliptic cross-section along the length of the coil. For this reason Curtis, Moon and Sparks[11] carried out the final drawing of the wire and the winding of the coil in one operation, the wire passing through specially selected dies straight on to the cylinder.

Difficulties may also arise in the fixing of the wire to the cylinder. This must be done in such a manner as not to affect the tension of the wire, and the terminal arrangements must be such as to make the number of turns and the total length of wire quite definite. Various fixing devices have been used: soldering to small brass plugs fitting into holes in the cylinder is probably as good as any, but there is always a danger of the turns near the ends of the helix being less satisfactory than those near the middle. In the new French[12] and American[11] standards these end turns are therefore not used in electrical measurements; current is passed into the helix via thin wires which are radially disposed and soldered to the helix at suitable points a few turns distant from the ends.

These details give some idea of the enormous amount of preliminary work required for an absolute electrical measurement of high precision. It must also be borne in mind that none of the materials available is strictly non-magnetic. In this respect fused quartz and marble are of equal merit, both being slightly diamagnetic: their permeability is less than that of a vacuum by about 10 parts per million. The correction arising on this account has been investigated. For the American coil on on a quartz cylinder, it is estimated as -3 parts per million, while for the French coil the correction is given as -8 parts per million.

The diameter of the helix is usually measured in some form of measuring machine against end standards of length, but the diameter of the French standard[12] recently constructed was deduced from a measurement of the length of wire wound on to the cylinder during the construction of the helix. The wire passed over a system of pulleys, one of which was suspended by a vertical loop of the wire itself. A weight attached to this pulley controlled the tension. Wire was unwound from the drum until the length of this loop was some 8 metres. The drum was then fixed and the cylinder rotated so as to wind on a definite number of turns (53) which required 16 metres of wire. The suspended pulley was thereby raised about 8 metres, and an accurate measurement of this height determines the length of wire used for the 53 turns. This process was repeated until the whole cylinder was wound. The method obviously involves many difficulties. A special pit had to be constructed to accommodate the long loop of wire; differences of temperature between the wire in the loop and that going on to the cylinder had to be allowed for; and a correction had

also to be made for the effects of differences of tension arising from friction at the pulleys. These differences of tension were estimated by an ingenious experiment in which a length of manganin wire, provided with leads so as to make it serve as a 4-terminal resistor, was included in the length of wire being wound. The resistance of such a manganin resistor is very easily measured with high precision and it is a comparatively simple matter to measure the change of resistance produced by known changes of tension. The wire was calibrated in this way, and thus became a very sensitive indicator of changes of tension. By measuring the changes in resistance which occurred during the various operations of winding, the changes in tension were deduced and the necessary corrections estimated. This correction amounted to about 10 parts per million, as did also that arising from temperature-differences. It is remarkable considering the difficulties of the measurements that the results of successive windings were reproducible to a few parts per million.

The tension of the wire must cause some compression of both wire and cylinder in the radial direction along their line of contact. Snow[14] has investigated this effect theoretically and shown how to calculate the effect of this elastic compression on the mean diameter from a knowledge of the tension and the elastic constants of the materials. The effect is, however, usually negligible.

The measurements of pitch must in all cases be made on the finished helix. Various methods have been used, e.g. a comparator at the International Bureau of Weights and Measures, the cylinder being horizontal, and a cathetometer at the National Bureau of Standards, the cylinder being vertical. These measurements are of course made by reference to line standards of length.

At the National Bureau of Standards, Washington, the Physikalisch-technische Reichsanstalt, Berlin, and the Laboratoire Central d'Électricité, Paris, single-layer coils such as are described above are used as primary standards of self-inductance. At the National Physical Laboratory, and the Electrotechnical Laboratory, Tokyo, primary standards of mutual inductance are employed. Standards of self-inductance are advantageous in that only one coil is required, and the calculation of the self-inductance of a regular helix with high accuracy is less laborious than that of a comparable mutual inductance. Standards of mutual inductance were first advocated by Albert Campbell on the grounds that by the use of a single primary coil and a subdivided secondary coil, almost any desired mutual inductance is easily obtained, the mutual inductance associated with the successive turns of the secondary coil being strictly additive. Moreover, Campbell showed that if the dimensions of the primary and secondary coils are suitably proportioned, the mutual inductance can be made independent of small variations in the size and position of the secondary coil to the first order of small quantities, and in such a case the dimensions of the primary coil only are required with high accuracy, and the difficulties of construction are therefore no greater than for a self-inductance standard: in each case only one single-layer helix constructed with the maximum of precision is required, and in both cases the accuracy of the standard is limited in practice by the irregularities of this helix. An important advantage of mutual inductance standards is that the relative uncertainties of mutual inductance due to

given irregularities of the helix may be made considerably less than the corresponding uncertainties of self-inductance. For the change in the self-inductance of a helix due to a displacement of one turn is proportional to the difference of the mutual inductances of that turn and the two end turns of the helix. Thus for turns near the ends of the helix this correction tends to become comparatively large, for it depends on the mutual inductance of turns which are very near together. Uncertainties in the relative positions of these turns, although very small in comparison with the total length of the helix, are not necessarily small in comparison with the distance between neighbouring turns, and the evaluation of the corrections becomes correspondingly difficult, for it may be impossible to neglect second order terms in the calculation. In a mutual inductance standard, however, when calculating the effect of the displacement of any one turn, we are concerned with the mutual inductance of this turn and the secondary coil, and the distance between these is always very large, compared with the displacement of the turn. The effect of the displacement is therefore always very small and easily estimated to a close approximation. When it is remembered that the uncertainties of such calculations are among the largest in modern absolute measurements, the importance of this advantage may be appreciated.

Formulae for the computation of the self-inductance of a helix, and the mutual inductance of a circle and a helix, have of course long been known, but among the recent work on the subject mention should be made of a new formula given by Snow[15] for the self-inductance of a helix, and also a formula[16] for calculating the effect of irregularities in pitch on the self-inductance of a helix. Astbury[17] has recently given a development of Viriamu Jones' formula for the mutual inductance of a helix and coaxial circle which considerably simplifies the numerical work.

The problem of establishing a primary standard of inductance is by no means completely solved when the coil itself has been constructed, its linear dimensions determined, and the value of inductance calculated, for until means of using the standard for electrical measurements have been devised, it can only be regarded as a geometrical standard. It is necessary to devise means of establishing relations of the type of (7) or (4) by means of the standard, or in more usual terms we must establish circuit arrangements such that the relation between the current I flowing through the standard and the quadrature component of the potential-difference V between its terminals is

$$V = jLwI \quad \text{or} \quad V = jMwI,$$

where L or M is the calculated value of inductance. In practice the terminal p.-d. depends also on the capacitance of the system: in other words, the coil when carrying current possesses an electric field as well as a magnetic field and both contribute to the terminal p.-d. The effects of capacitance increase rapidly with the frequency of the alternating current used and therefore primary standards are used with current of low frequency only. Standards of mutual inductance are more troublesome in this respect than those of self-inductance, since the distribution of capacitance in the system is more complicated when two coils are present. At the National Physical Laboratory, a frequency of 10 c./sec. is used for work of this kind.

Astbury[18] in a recent paper, has given a detailed examination of the effects of capacitance in measurements with a mutual inductance standard of the Campbell form, and has shown that under suitable conditions comparative measurements can be made with an accuracy approaching 1 part per million. Such measurements are greatly facilitated by a device by means of which the standard has been made continuously adjustable over a small range. Two series of extra turns have been added to the secondary coil. From one of them an integral number of turns may be selected by means of a switch and connected in series with the main secondary coil. The remaining extra turns are connected in parallel with a small potentiometer, the potential circuit of which is also connected in series with the main secondary coil. The potentiometer thus subdivides the p.-d. across these last turns, and therefore in effect allows any desired fraction of a turn to be added to the secondary coil.

§ 9. ABSOLUTE DETERMINATIONS OF THE OHM

There are many historical absolute determinations of the ohm, but the first measurement of high precision (by modern standards) was that carried out by F. E. Smith[19] in 1913 using the Lorenz apparatus of the National Physical Laboratory. This is the apparatus required to establish the first experimental law (4) in Norman Campbell's scheme. Essentially, this apparatus consists of two primary standards of mutual inductance designed in accordance with Albert Campbell's principle, mentioned above, but the secondary circuits take the form of metal discs which can be rotated at uniform angular velocity. The rotating discs are equivalent to, and indeed actually carry, rotating radial conductors, and in the experiments the e.m.f. generated in these conductors is balanced by the potential difference across a resistor which is in series with the primary coils of the mutual inductors. This state of balance is independent of the value of the current and therefore serves to establish the required relation between the calculated values of mutual inductance and the speed of rotation of the discs or secondary conductors. The difficulties of this experiment are almost entirely of a practical character. The e.m.f. available for measurement is very small, and in order to maintain contact with the revolving conductors a rubbing contact must be employed. By the use of special brush-contacts constantly lubricated with benzene undesired thermo- and contact-e.m.fs. were reduced to a minimum. Further, the two inductors were so arranged that these unwanted e.m.fs. (and also those due to the earth's magnetic field), were in opposition while the e.m.fs. of mutual induction were added, and in this way a sensitivity of the order of 10 parts per million was achieved in spite of the fact that the e.m.f. available for measurement was only of the order of a few millivolts. Nevertheless, the accuracy of the experiment is always limited by the degree of steadiness and sensitivity attainable in the electrical balance, and it is always necessary to take the mean of several observations in order to derive a measurement.

Smith's determination has in the last few years been repeated by P. Vigoureux[20]. The apparatus employed was the same as was used by Smith, but the linear dimensions of the coils were re-determined, and certain improvements were made in the

auxiliary equipment used for adjusting the positions of the coils and determining their distance apart. The result obtained was

$$1 \text{ international ohm (N.P.L.)} = 1{\cdot}000\ 50 \text{ ohm},$$

while Smith had obtained the value 1·000 52 for the corresponding quantity in 1913. The estimated uncertainty of each measurement was 20 parts per million apart from uncertainties in the value assumed for the international ohm, so that no significance can be attached to the difference between the two results.

Smith's determination of 1913 was closely followed by one of comparable precision carried out by Grüneisen and Giebe[21] at the Physikalisch-technische Reichsanstalt in 1914 but not published until 1920. The latter determination was the first one of high precision in which alternating currents were used, although the principle of employing alternating currents for this purpose had already been explored by Albert Campbell[22] who had published two methods, one of which is very similar in principle to that employed by Grüneisen and Giebe and others. It is unnecessary to enter into details here but the principle is as follows. A primary standard of self or mutual inductance is constructed, and the capacitance of a condenser is then measured in terms of this standard by means of an alternating current bridge, such as Maxwell's or Carey-Foster's. This condenser is then balanced against a standard resistor in Maxwell's commutator bridge and a relation thereby established between resistance capacitance, and speed of commutation. Thus resistance is measured in terms of time and capacitance, which is measured in terms of inductance, which is calculated from linear dimensions. The method involves the assumption that the capacitance of the condenser under the conditions imposed by the commutator-bridge is the same as that under those of the alternating current bridge, which means in practice that the condenser must be sensibly free from dielectric absorption so that its capacitance is independent of the rate of charge and discharge. Albert Campbell had only mica condensers and although he made an estimate of the correction due to absorption, the uncertainty of this correction limited his accuracy to about 2 parts in 10^4. Grüneisen and Giebe used standard air condensers, a unique series of these instruments[23] having been constructed at the P.T.R. Measurements made with different speeds of commutation showed that the effects of dielectric absorption were negligible in these condensers. The value obtained by Grüneisen and Giebe for the ratio of the international ohm to the ohm was 1·000 51 with an estimated uncertainty of ± 3 parts in 10^5. It was subsequently found that the resistors and mercury tubes used for the derivation of the international ohm had changed since their last calibration, and it was estimated[24] that the value for the above ratio should be 1·000 49.

It was on the evidence provided by the determinations of Smith and Grüneisen and Giebe that the decision to abandon the international ohm in favour of the ohm was reached, for the difference between the results of these two methods was no greater than the uncertainty in the value of the international ohm, and there seemed good reason to hope that the uncertainty might be still further reduced if absolute determinations of high precision were made by as many national laboratories as

possible. No less than six such determinations were put in hand, and of these four have now been published.

The first of these was that by Curtis, Moon and Sparks[11] carried out at the National Bureau of Standards, Washington. The method was exactly the same as that of Grüneisen and Giebe, but the apparatus employed embodied many refinements. The vast amount of detailed work on the primary standards of inductance has already been mentioned (p. 312). An exhaustive study was also made of the Maxwell commutator bridge;* the commutator consisted of a tuning fork interrupter, the frequency of which was controlled by a piezo-electric oscillator, and was therefore constant and of nominal value to 1 part per million. The contacts of the interrupter were mounted on the limbs of the tuning fork, and in some cases were contained within evacuated glass tubes. Unfortunately, these refinements did not have the desired effect of reducing the discrepancy between the various values for the ohm. The value obtained for the ratio of the international ohm to the ohm was 1·000 45. The investigation showed, however, that the major uncertainty arises from irregularities in the pitch of the winding, and suggested that the agreement between the values of Smith and Grüneisen and Giebe was to some extent fortuitous. Subsequently a standard inductor of more uniform pitch was constructed by lapping a uniform helical groove in a glass cylinder and winding the wire in the groove. Measurements made with this inductor gave for the above ratio a value of 1·000 48[25]. The authors consider that when due weight is attached to their original value, their final result should be taken as 1·000 47.

The National Bureau of Standards determination of 1936 was quickly followed by two new determinations at the National Physical Laboratory. One of these, Vigoureux's determination by the Smith-Lorenz apparatus, has already been discussed: the other was a determination by Hartshorn and Astbury[26], which consisted of an elaboration of a new alternating current method outlined by Albert Campbell[27] in 1925. It has been mentioned that one of the sources of error in Campbell's earlier determination was the difference in behaviour of the circuit components and especially the condenser with respect to the sine-wave voltage used in the Carey Foster bridge and the interrupted voltage used in the Maxwell bridge. Campbell's new method completely avoids this difficulty for it requires sine-wave voltage only, and dispenses with the condenser. The object of any alternating current method is to compare an e.m.f. of self or mutual inductance ($j\omega MI$ or $j\omega LI$ in the usual notation) with the potential difference RI across the terminals of a resistor carrying the same current I as the inductor. Practical difficulties arise because the two p.-ds. are in quadrature, but if two mutual inductors are placed in cascade, the final secondary p.-d. may be brought approximately into phase with the first primary current, and may therefore be balanced against the p.-d. across a resistor through which this current also flows. Campbell's new circuit was essentially an alternating current potentiometer circuit based on this principle. In practice the quadrature component of voltage cannot be completely

* Cf. *Progress Reports*, volume 1, 341 (1934).

avoided, but it may be balanced by an auxiliary adjustment. The resistance is measured by the condition of balance for the in-phase p.-d. which is of the form

$$R_1 R_2 = M_1 M_2 \omega^2, \qquad \ldots\ldots(9)$$

where R_1 and R_2 are two resistances in the circuit and $M_1 M_2$ the two mutual inductors. (See figure 1.) The ratio R_1/R_2 is easily measured, and this ratio and the relation (9) give the values of both R_1 and R_2 in terms of $M_1 M_2$ and the frequency, and the values of these quantities may be deduced from measurements of length and time respectively.

Hartshorn and Astbury[26] made a determination by this method, after examining the various sources of error in great detail, and their experiments show that the method is capable of very high precision. The average deviation of their results from the mean value amounted to only 3 parts per million and this in spite of the fact that some of the measurements were made with different frequencies, and different values of resistance. The total estimated uncertainty of their result was $\pm$ 15 parts per million of which $\pm$ 10 parts per million was in the value of the primary standard of inductance. It is not practicable to use calculable inductance standards in the Campbell circuit itself for, as previously mentioned, such standards can only be used at very low frequencies, and these are not very suitable for the Campbell circuit. The plan adopted was to construct two auxiliary standards M_1 and M_2 of nearly the same value as the primary standard but suitable for use at higher (100 c./sec.) as well as low frequencies. The method then consists of the following measurements (*a*) comparisons of M_1 and M_2 with the primary standard, (*b*) evaluation of $R_1 R_2 = M_1 M_2 \omega^2$ by the Campbell circuit, (*c*) measurement of the ratio R_1/R_2. A sensitivity of 1 part per million is easily obtained for all these measurements, for (*a*) and (*b*) are carried out with alternating current, and it is therefore possible to use a thermionic amplifier in conjunction with the detector of balance, while (*c*) is a simple d.-c. resistance comparison. The frequency is easily maintained constant and measured to better than 1 part per million by use of a modern standard of frequency of either the piezo-electric or tuning fork types. Nevertheless it is by no means a simple matter to obtain an accuracy of a few parts per million in the complete system of measurements, for it is impossible in practice to realize the ideal circuit postulated by theory. In any actual circuit, inductance both self and mutual are distributed throughout the entire system, and not localized in particular components. Moreover, capacitance is also distributed throughout the entire system, and the distribution of both inductance and capacitance is apt to be altered by such slight changes as are necessary in order to make the various measurements possible. These sources of error were dealt with by the technique of modern alternating-current bridge work, e.g. earth-capacitance by the Wagner earth-connexion, stray mutual inductance by the use of coaxial or twisted leads, twisted leads where a symmetrical distribution of capacitance is desired, and coaxial leads where electrostatic screening is desired, and so on. Harmonics in the current supply, which were found to be troublesome, were eliminated by the use of filter-circuits. An accuracy of a few parts per million can only be obtained by close attention to the most minute

details in the construction and arrangement of the circuit, but when once the equipment has been constructed, the necessary observations can be made quickly and easily so that the apparatus serves as a convenient primary standard equipment.

The most recent precision determination of the ohm was that carried out by Jouaust, Picard and Héron[12] at the Laboratoire Central d'Électricité in Paris. The method is a modification of Campbell's last method due to Picard[28]. The nature of the modification will be understood by reference to figures 1 and 2. Campbell's equation $R_1R_2=M_1M_2\omega^2$ becomes $R_1R_2=M_1L\omega^2$, where L is the total inductance of mesh (2) of Picard's circuit. A primary standard of self-inductance was employed, and L was measured in terms of this standard by substitution in a simple comparison bridge of the Wheatstone type. The primary of the mutual inductor was of self inductance L' which was nearly equal to both L and the self-inductance of the primary standard. The mutual inductance M was compared with the self-inductance L_0 of the primary coil, by means of the Maxwell-Rowland bridge, and L_0 was compared with the primary standard by substitution. Picard's circuit has the advantage that both the resistances R_1 and R_2 are values of non-inductive resistors,

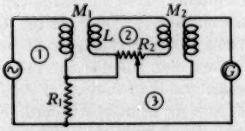


Figure 1. Campbell's circuit.

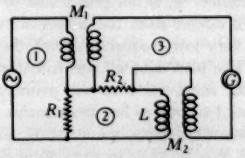


Figure 2. Picard's modification.

which may therefore be made practically independent of frequency and temperature and very easily compared and measured in international ohms. In Campbell's circuit R_2 includes the resistance of inductive coils, and is therefore not quite so easily measured; the temperature of the coils must be accurately controlled, and their frequency corrections determined. In other respects Campbell's method is the simpler, since the number of measurements required for a determination is smaller.

The technique employed for these two methods is essentially the same in spite of differences in points of detail. The value obtained by Jouaust, Picard, and Héron for the ratio of the international and absolute ohms was 1·000 52 with an estimated uncertainty of perhaps several parts in 10^5.

The above investigations have shown that absolute measurements of resistance by at least three distinct methods agree to nearly 1 part in 10^5 which is probably the maximum accuracy to be expected with present-day technique. If our object were merely to define a unit of resistance on the absolute principle with the highest possible precision, it would be necessary to do as Norman Campbell has suggested and define the ohm in terms of one selected method, after which all other methods would have no further interest in this particular connexion. However, the agreement between different methods probably constitutes the most accurate experimental evidence available in support of the equations of electromagnetic theory on

which practice so largely depends. The measure of agreement obtained by the widest possible variety of methods is therefore a matter of some importance, and investigations of new methods or improved forms of old methods are of value even though their accuracy may prove to be inferior to that already obtained by other methods. In this connexion several other recent investigations on the absolute measurement of resistance are of interest. Nettleton and Llewellyn have made measurements by two methods based on the old method of damping devised by Weber in 1851. In this method a magnet of moment m is suspended at the centre of a coil of wire, and its period of vibration T and logarithmic decrement λ observed, first with the coil open ($T_0\lambda_0$) and then with the coil closed ($T\lambda$). The resistance (R) of the coil is then obtained from the relation

$$R = mG^2(1+A_1)^2/4I\,(\lambda/T - \lambda_0/T_0) + 4\lambda L/T,$$

where G is the galvanometer constant of the coil, L its self inductance, I the moment of inertia of the magnet and A_1 a correcting factor depending on the angle of oscillation. Mere inspection of this formula is sufficient to show that the probable errors of the method would be great. In their first experiment Nettleton and Llewellyn[29] succeeded in replacing the quantity $m^2G^2(1+A_1^2)/4I$ in this equation by other quantities more easily measured. They used a coil which could be rotated, its terminals being connected to a galvanometer through contacts which are closed during a small fraction of each revolution. An e.m.f. is induced in this coil as it cuts the flux of the magnet and the galvanometer, therefore, tends to deflect. The deflection is annulled by the flux of a compensating coil carrying an adjustable current i_m. The value of the current i_m and the mutual inductance M between the rotating and compensating coils at the instants of make and break are obviously a function of $mG(1+A_1)$. The rotating coil and magnet were also used as a sine galvanometer and the current i_θ required to produce the standard deflection noted. It was shown that the factor $m^2G^2(1+A_1)^2/4I$ is then replaced by

$$M\,(i_m/i_\theta)\,(\pi^2+\lambda_0^2)/T_0^2,$$

so that the only quantities required in addition to the vibration constants are the mutual inductance M, and the current ratio i_m/i_θ which may be accurately measured by a potentiometer method. In spite of these improvements it appears that the accuracy of the method is limited to about 1 part in 10^3, for the equation of motion of the oscillating magnet is only applicable to oscillations of small amplitude. This limitation is overcome to a considerable extent in a later investigation[30] in which the magnet is replaced by an oscillating coil. No rotating coil is then necessary. The apparatus consists merely of an electrodynamometer with concentric circular coils of massive construction. It is shown that if the ratio of the radii of the two coils has a certain value (0·506078 for concentric circles), their mutual inductance is so accurately proportional to angular displacement of one coil from the conjugate position, that the deviation from the linear law is less than 5 parts per million for a deflection as large as 7°. Thus the deflecting couple of such a dynamometer is accurately proportional to the product of the currents in the two coils and the angular

deflection, and if the suspended coil is closed and allowed to oscillate, the damping couple is accurately proportional to the velocity, and for oscillations of considerable amplitude there is remarkably close approximation to true damped simple harmonic motion. It is shown that the resistance R of the oscillator circuit is given by

$$R = M \frac{i_f}{i_m} \Big/ \left[\frac{\lambda T}{\pi^2 + \lambda^2} - \frac{\lambda_0 T_0}{\pi^2 + \lambda_0^2} \right],$$

where $\lambda_0 T_0 \lambda T$ are as before values of logarithmic decrement and period with the coil open and closed respectively, i_f is the current in the fixed coils, and M is the mutual inductance between the coils in the position of steady deflection obtained when a current i_m flows in the moving coil. With this apparatus an accuracy of a few parts in 10^4 was obtained. The form of the governing equation and the number of measurements involved suggest that it would be difficult to obtain an accuracy much greater than this by the method of damped oscillations. However, by an application of the principle of the rotating coil to a coil-system similar to that of the above-mentioned dynamometer, Nettleton and Balls[31] have devised a method which appears to be capable of considerably better accuracy. A constant current is passed through the resistor to be standardized, and the fixed coil system. The other coil rotates at constant angular velocity and the e.m.f. induced in this coil is balanced against the p.-d. across the resistor, a sensitive galvanometer being used as detector of balance. As in the previous method, the coils are so proportioned that their mutual inductance is very nearly proportional to angular displacement from the conjugate position over a considerable range. Within this range the e.m.f. induced in the rotating coil is approximately constant, so that by using a commutator which closes the balancing circuit during this portion of the rotation only, the condition of balance may be determined with high precision. The equation of balance takes the form

$$R = \omega . M/\theta,$$

which determines the value of the resistance R in terms of the angular velocity ω and the mean change of mutual inductance per unit angle of rotation (M/θ) throughout that portion of each rotation during which the commutator closes the galvanometer circuit. The advantage of this method over the Lorenz method is that the e.m.f. to be balanced may be made comparatively large so that a resistance of the order of 1 ohm may be measured and also greater sensitivity may be obtained. However, the design of the coil-system does not permit the calculation of M/θ with high precision, so that the method is less direct than the Lorenz method. Nettleton and Balls determine corresponding values of both M and θ for settings of the coils corresponding to the limits of the commutator by reference to a separate standard of mutual inductance. The angles are measured electrically by the use of an additional coil rigidly attached to the moving-coil system; and an additional pair of fixed coils in the Helmholtz position. It is shown that the mutual inductance between a solenoid of certain dimensions and a pair of large Helmholtz cells between which it rotates can be made with great accuracy proportional to the sine of the angle of displacement from the conjugate positions. Thus any angular setting θ for such a

coil may be determined from a measurement of the ratio of the mutual inductance at this setting to the maximum value (M_θ/M_{max}). Thus the constant M/θ for the apparatus may be deduced from measurements of mutual inductance only, and provided that the apparatus is designed so that M/θ is practically independent of θ for settings near the limits of the commutator, this constant of the apparatus should be measurable with high accuracy. The experiments made show that with a resistance of the order of 1 ohm ample sensitivity is easily obtained; an accuracy of about 1 part in 10^4 can be obtained without great elaboration.

It is to be observed that the methods which have proved to be capable of the highest precision are null-methods in which a condition of balance is observed. Methods have also been devised which depend on the condition of equality of two alternating currents. Such methods are very simple theoretically but in practice it is more difficult to observe with precision a condition of equal currents than one of

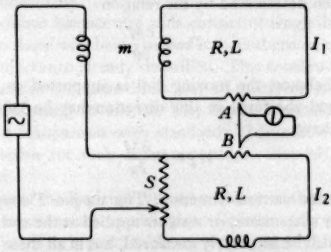


Figure 3. Campbell's equal current method.

zero current. The presence of harmonics in the two currents also invariably causes considerable difficulty. The simplest of such methods is one due to Nettleton and Balls[32]. Alternating current of frequency $\omega/2\pi$ is passed through the primary coil of a mutual inductor. The resistance R of the secondary circuit is then varied until the secondary induced current is equal to the primary current. The resistance R may then be obtained from the relation

$$R = \omega\sqrt{(M^2 - L^2)},$$

where M is the value of the mutual inductance, and L is the self-inductance of the secondary circuit. A. Campbell[33] has also devised a circuit of this type, figure 3. The two loops must be adjusted to have equal total resistance R and inductance L. The mutual inductance M or the resistance S are then adjusted until the root-mean-square values of the currents I_1 and I_2 are equal. These two currents are then in quadrature, and we have

$$S = M\omega.$$

Campbell observes the condition of equality of currents by means of differential thermo-junctions and heaters A and B in figure 3. The use of this circuit as a device for the production of equal currents in quadrature (phase-splitting) has been patented.

§ 10. ABSOLUTE DETERMINATION OF THE AMPERE

The methods which have been devised for the absolute measurement of current are far less varied than those for the determination of the ohm. In recent years work on the subject is almost exclusively confined to two methods, those of the electrodynamometer and the current balance, and of these only the current balance has yielded results of high precision. In both methods the current to be measured is passed through two circuits in series, one of them being fixed and the other supported so as to be capable of moving freely either about or along a fixed axis. The passage of the current gives rise to mechanical forces between the two circuits and the movable circuit therefore tends to become displaced. In the electrodynamometer the movement about an axis is opposed by a measurable mechanical torque, T, which when equilibrium is established is equal to that due to the current I. The value of I is then determined by the relation

$$T = I^2 \frac{\partial M}{\partial \theta}.$$

In the current balance, the moving coil is supported so that its direction of motion is vertical, and the force in this direction may be balanced by a weight W and in this case we have

$$Wg = I^2 \frac{\partial M}{\partial z},$$

where z represents the vertical direction. The torque T may be produced by a torsion wire, a bifilar suspension, or weights applied at the end of a horizontal arm, the length of which can be accurately measured, but in all these cases the determination of T involves a measurement of mass and in addition measurements of other quantities such as lengths and periods. It follows that the law of the current balance, which is free from these additional quantities is the simpler, and it is therefore not surprising that the current balance has proved to be capable of the higher precision. The quantities $\partial M/\partial \theta$ and $\partial M/\partial z$ are of course both determined from the linear dimensions of the two circuits.

The problem of the absolute determination of the ampere therefore reduces to that of the construction of two circuits of non-magnetic material, and of such a form that $\partial M/\partial z$ may be calculated or otherwise determined with the highest accuracy, and the mounting of one of the circuits so that it is capable of free motion in the vertical direction. An ordinary laboratory balance of high sensitivity obviously supplies the form of mounting required, if the movable coil takes the place of one of the usual scale pans. Only two such instruments of high precision have so far been constructed: the Ayrton-Jones balance of the National Physical Laboratory and the Rayleigh balance of the National Bureau of Standards, Washington. Absolute measurements of current by means of both these instruments were first made some thirty years ago and these have recently been repeated.

It is obvious that the value of $\partial M/\partial z$ for the two coils will only be determinable with high accuracy if their linear dimensions are well defined and constant. The

necessary rigidity can only be secured by a somewhat massive construction, and the weight of the coils is therefore considerable: indeed the force due to the current is usually quite small in comparison, for it is impracticable to use either very large currents or a very large number of turns of wire, the former because of disturbances due to the heat generated and the latter because of the difficulty of securing well-defined linear dimensions. It therefore follows that the balance itself must be capable of extraordinary sensitivity, e.g. in the National Physical Laboratory instrument the moving system weighs over 14 kg. but it will measure differences in force to the nearest 0·01 mg. weight, which corresponds to a sensitivity considerably better than 1 part in 10^8 in the total force operative. A similar sensitivity is obtained with the American balance. In each case the observations are made by means of a mirror mounted on the beam of the balance, the observer working with a scale-distance of several metres and observing either by means of a telescope or the usual lamp and scale arrangement.

In the Ayrton-Jones balance the coils consist of single-layer solenoids wound on marble cylinders, bare wire being wound in helical grooves on the cylinders as in the standards of inductance already described. This construction originally due to Viriamu Jones provides the most direct way of satisfying the conditions required for the accurate evaluation of $\partial M/\partial z$. The linear dimensions of the helices are measured by direct comparison with standards of length, every turn of the coils being readily accessible for such measurements, so that $\partial M/\partial z$ can be calculated with great accuracy.

Current balances of the Rayleigh type are so designed that such detailed measurements of linear dimensions are unnecessary. Circular coils with a cross-section of winding that is as small as is practicable are employed, and they are disposed so as to be coaxial, the axis being vertical, and their axial distance being such that the vertical force between them has a maximum value. The value of $\partial M/\partial z$ for two such coils is obviously a function of their radii r_1 and r_2 and their axial distance z. At a certain distance z_m this function has a maximum value, which depends on r_1 and r_2 only; and since $\partial M/\partial z$ is of zero dimensions in length, this maximum value must be a function of the ratio r_1/r_2 only. This ratio may be determined by a purely electrical method. The two coils are removed from the balance and mounted so as to be concentric and coplanar and a small magnet is suspended at their centre so that it is deflected by a current passing through either coil. Currents are then passed through both coils in directions such that their effects on the magnet are opposite, and these currents are adjusted until their combined effect on the magnet is reduced to zero. The ratio of the two currents multiplied by the ratio of the numbers of turns of the coils then gives the ratio of their mean radii. In the balance the fixed coil is duplicated, the suspended coil being placed midway between two similar fixed ones, through which current passes in opposite directions. The force to be measured is thereby doubled. It is also important to notice that for values of z near z_m the value of $\partial M/\partial z$ is independent of small errors in the value of z.

The Ayrton-Jones balance of the National Physical Laboratory was constructed nearly forty years ago, and was first used by Ayrton, Mather and Smith[34] for a

determination of the e.m.f. of the Weston standard cell, which was published in 1908. The weighing mechanism is built on conventional lines but is unusually massive and is supported on a massive pedestal of phosphor-bronze. A helix wound on a marble cylinder is supported from each end of the beam, and a corresponding pair of fixed coils rest on the supporting pedestal. By means of adjusting screws the position of each fixed coil may be adjusted in the vertical direction and in two horizontal directions at right angles. Each fixed coil consists of two helices separated by a short distance, the current passing in opposite directions round the two helices, so that when the fixed and moving coils are coaxial and their mid-planes coincide, the forces on the moving coil due to the two parts of the fixed coil are in the same direction. This feature of the design corresponds to the two fixed coils of the Rayleigh balance. The duplication of the complete fixed and moving coil systems again doubles the force to be measured, and also preserves symmetry within the balance case even when current is flowing and heat is therefore being generated. The current passes from the fixed to the moving coils by way of two groups of 40 very fine silver wires in parallel, which are sufficiently far from the coils to have no appreciable effect on them. They reduce the sensitivity of the balance by only 6 per cent. The remainder of the wiring is done with coaxial cable which of course has no external magnetic effect. The linear dimensions of the coils are determined before they are mounted on the balance. Afterwards the adjustment to the symmetrical position is made electrically, the three adjustments of each fixed coil being operated in turn until the position of maximum or minimum force is found for each coil. For these measurements the current passes in directions such that the forces acting on each suspended coil are in the same direction, so that the resultant turning moment on the beam is nearly zero, and no manipulation of weights other than the rider is necessary. The position of the coils having been adjusted, current is passed through them and a standard four-terminal resistor in series, the connexions of the coils being such that the turning moments of the two systems are in the same direction. The current is then varied until the p.-d. across the resistor R exactly balances the e.m.f. E of a standard cell. The current then has the value E/R and is maintained constant at this value while the weighings are performed, the constancy being judged by the balance of the circuit which includes the standard cell. Readings are taken before and after reversing the current in the fixed coils so as to allow for the effect of the earth's field, and also before and after reversing the current in both fixed and suspended coils of one system, which reverses the secondary forces of the system (i.e. forces between the suspended coil of one system and the fixed coil of the other), but has no effect on the direct forces. Thus four readings are required, the mean of the four giving the direct force to which the basic equation applies. The value of the current may then be calculated, and equating this to E/R, a value for E the e.m.f. of the standard cell is obtained, in terms of the ampere and the unit of R. The same combination of standard cell and resistor may then be used to control the current passing through a silver voltameter, and E may be determined in terms of the international ampere and R. Comparison of the two results then gives the relation between the international and absolute amperes.

Absolute measurements of current were made with the N.P.L. balance by Ayrton, Mather and Smith[34] in 1905–7 and again by Smith[39] in 1910. These were the first absolute measurements to be made with present day precision. Some years later it was observed[37] that the original coils were deteriorating. The gap between adjacent turns of the coils amounted to only 0·15 mm. and in order to improve the insulation by excluding dust particles the windings had been coated with paraffin wax. In the course of time the wax became oxidized and the acids formed not only greatly reduced the insulation resistance but also corroded the copper wire. It therefore became necessary to construct new coils. The pitch of the windings was increased so that the gap between adjacent turns was trebled, and no wax was then necessary. The number of turns and therefore the force to be measured were considerably decreased by this procedure, but the sensitivity of the balance was still adequate. Silver wire, gilded in order to prevent tarnishing, was used for the windings, but in practice this has been found to tarnish more than bare copper. The reconstructed balance was recently used for a new determination of the ampere by P. Vigoureux[36]. The technique of the measurements was essentially the same as that employed in the earlier measurements but the various sources of error, particularly those involved in the calculation of $\partial M/\partial z$ were investigated in even greater detail than before. The number of measurements of linear dimensions was greatly increased, and it was shown that the result obtained by taking the mean of only half the measurements differed by only 2 parts per million from that obtained by taking the mean of all. The difference between the forces exerted by the two systems of coils was also calculated and measured experimentally, and determinations were made using each system separately. The linear dimensions were also checked while current was flowing through the coils, and calculations were made of the effect of slight departures from the coaxial condition. The general consistency of all these measurements showed that the calculated value of $\partial M/\partial z$ was probably correct to a few parts per million. The effect of a relative rotation of fixed and suspended coils about their axis was also investigated theoretically and practically, and it was shown that although the force between the helices varies slightly with azimuth, it is not difficult to find the setting at which the value of the force is equal to the mean value for all settings, which is given by Jones' original formula for $\partial M/\partial z$.

The final result may be expressed in the form:

e.m.f. of Weston neutral cell at 20° C.

$$= 1{\cdot}01816 \text{ international ohm} \times \text{absolute amperes.}$$

Measurements made with silver voltameters in 1931 had given

e.m.f. of Weston neutral cell at 20° C.

$$= 1{\cdot}01830 \text{ international ohm} \times \text{international amperes}$$

from which it is deduced that

$$1 \text{ ampere} = 1{\cdot}00014 \text{ international amperes.}$$

This value for the ratio of the ampere to the international ampere is identical with that obtained by F. E. Smith in 1910, while his value for the e.m.f. of the Weston

cell in international ohms × absolute amperes was only 20 parts per million greater. It therefore appears to be probable that over a period of twenty-five years, standard cells, and also the values of current determined by the silver voltameter and the Ayrton-Jones balance have all been reproduced with an accuracy of 2 parts in 100,000.

In 1910 McCollum[38] suggested that in measurements with an electrodynamometer the torque due to the current might with advantage be deduced from measurements of the periods of oscillation of the moving system with and without current flowing through the coils. F. E. Smith[39] in 1912 showed that the same principle could be applied to the National Physical Laboratory current balance. It is merely necessary to connect the coils so that the forces due to the current cause no change in the position of equilibrium of the beam but a large change in the restoring couple. The moment of inertia of the moving system must also be determined, and its determination is not quite so simple as appears at first sight, since the beam is apt to bend and therefore to change its moment of inertia with change of load. However, a scheme was devised whereby a small weight on the moving system could be displaced by a measured amount in such a manner as to cause a strictly calculable addition to the moment of inertia. Measurements of the time of oscillation before and after this change then provide all the data required for the determination of the moment of inertia in the original working condition. When the balance was reconstructed the fittings required for these measurements were added, and the dimensions of the new coils were chosen so as to make the restoring force for a small axial displacement δz nearly proportional to δz, the terms immediately following the first in the series representing this force being reduced to zero. In practice, however, the accurate measurement of the time of oscillation with current flowing proved to be difficult owing to disturbances produced by the heating effect of the current. The periodic time was found to vary considerably with temperature, so that measurements of short periods of time to a very small fraction of a second are necessary, which is difficult in view of the necessarily low velocity of the beam even when passing through the the rest point. At the moment, therefore, the accuracy of the pendulum-balance method must be considered to be of a lower order than that of the current weigher.

The precision balance of the National Bureau of Standards was devised by Rosa Dorsey and Miller[40], and used by them for a series of absolute measurements published in 1911. Quite recently it has been used for a new series of measurements by H. L. and R. W. Curtis[41]. The balance being of the Rayleigh type, access to the individual turns of the windings for the purposes of determining linear dimensions is not required, and it is therefore possible to use multi-layer coils wound with insulated wire, on a metal former. With this construction it is possible to make the force to be measured comparatively large without at the same time making the suspended coil excessively heavy. The heating effect of the current is always a difficulty but with coils on metal formers water-cooling becomes possible. The balance of Rosa Dorsey and Miller was constructed on these lines. Three or four sets of fixed and moving coils were constructed but only one set was used at a time. The balance therefore lacked the symmetry of the Ayrton-Jones instrument but the

complication of secondary as well as direct forces was avoided, and the manipulation was correspondingly simpler. The fixed coils were of large diameter (up to 50 cm.) and consisted of 400 to 650 turns of enamel covered wire wound in a channel of square section cut in a brass former. Excessive heating was prevented by circulating water in a second channel in the former. The suspended coils, which were also wound on brass formers, were of about 20 cm. diameter and had rather less than 100 turns. Direct water-cooling of the suspended coil is of course impracticable, but the coil was surrounded by a fixed water-cooled shield, which removed much of the heat radiated from the moving coil and also served to minimize convection currents. Moreover, the coil-system was not contained within the case of the balance itself, but in a second chamber immediately below it, the brass tube by which the moving coil was suspended from the beam passing through a small hole in the base of the balance-case. The balance was operated in the stable state which was attained after current had passed through the system for some hours. In this respect it differs from the Ayrton-Jones balance which must be operated immediately after switching on the current and before the disturbances due to the heating effect become serious.

For the investigation published in 1934, H. L. and R. W. Curtis[41] used the coils constructed by Rosa Dorsey and Miller, although many of the details such as the water-cooling system were modified. The ratio of the mean radii of the various pairs of coils was measured by the magnetometer method already mentioned. For this purpose the coils must be mounted so as to be accurately concentric and coplanar, with their common plane vertical and approximately in the meridian, and it is obvious that the attainment of this condition with the necessary accuracy will involve many fine adjustments. For these measurements the coils were mounted on a vertical stand constructed by cementing together large sheets of plate glass, the coils being held by gimbal systems attached to the stand, and provided with adjusting screws, by means of which the large coil could be displaced bodily in three directions at right angles, and rotated about two rectangular axes, while the small coil could be displaced in one direction and rotated about one axis. The suspended magnet was of cobalt steel 1·5 to 3 mm. long, attached to a glass rod which was suspended by a quartz fibre, and carried also the wing of an insect to serve as a damping vane. Great care was taken to select non-magnetic material for the whole of this apparatus. The case containing the suspended magnet could also be displaced bodily, with respect to the plate-glass stand. The various adjustments were operated in turn until the required arrangement was reached by successive approximation. Each adjustment was made electrically, that is to say, currents were passed through the two coils in opposite directions and their ratios adjusted until the magnet showed no deflection: one coil was then slightly displaced and the measurement repeated, and in this way the ratio of the currents was determined as a function of the adjustment under investigation. In each case this function has a maximum or minimum value for the correct setting, which is therefore given by the observations. The actual current-ratio was measured by an application of the Wheatstone bridge principle. The two coils formed two adjacent arms of the bridge, the other arms

consisting of resistors, one of which is capable of very fine adjustment. When the bridge is balanced the ratio of the currents in the two pairs of adjacent arms is equal to the ratio of their resistances, and in this way the ratio of the two currents is given by a resistance ratio which is easily measured by standard methods. It is unnecessary to enter into further details. It will be obvious that with currents as large as 1 ampere great sensitivity can be obtained though the whole sequence of observations is somewhat tedious. Measurements having been made in this way for both fixed coils, they are mounted on the balance and their positions adjusted until the conditions required by the theory of the Rayleigh balance are satisfied. The levelling of the coils is carried out mechanically by the use of a sensitive spirit level, but the other measurements are made electrically. The fixed coils are adjusted so as to be coaxial and at that distance apart which corresponds to maximum force by measurements of the difference between the forces exerted by the two fixed coils on the moving coil midway between them. For the correct setting this force is independent of small displacements of either of the fixed coils in both the vertical and horizontal directions. This position is found by trial, the position of one of the fixed coils relative to the other being adjustable in the vertical and two horizontal directions at right angles. The moving coil is then adjusted to be coaxial with the fixed coils and midway between them by measurements of the sum of the forces between the moving coil and the two fixed coils, the fixed system as a whole being provided with adjustments for this purpose. The required position is that for which the force has a maximum value with respect to vertical displacements and a minimum value with respect to horizontal displacements, and this position is also found by trial.

The coils then being in the correct position, an actual weighing is made. Current is passed through the coils until a steady state is reached and the rest point is then observed. The current through the fixed coils is then reversed and simultaneously the appropriate weight added to the balance and the rest point again observed, the whole operation being performed without arresting the balance. The two readings are repeated until the required degree of precision has been obtained. The adjustment of the current to a constant value and the recording of the result in terms of the e.m.f. of the Weston cell are carried out in the way already described for the National Physical Laboratory balance.

The ideal Rayleigh balance has coils of negligibly small cross-section, and in practice it is necessary to apply corrections to the simple formula for the force between circular linear currents, in order to allow for the finite cross-section of the coils used. Such corrections were discussed by Rosa Dorsey and Miller in their original paper, but Snow[42] has recently given formulae which are applicable with a closer approximation to the practical case of a Rayleigh balance with coils of rectangular cross-section. The linear dimensions of the cross-section are of course small compared with the mean radii of the coils and Snow neglects only terms of the sixth and higher orders in the ratios of these quantities. Not only is the magnitude of the force between the coils dependent on their finite cross-section, but it also depends on the distribution of current over this section, i.e. on the fact that the current is concentrated in round wires separated by insulating material, and that even over

the cross-section of the wire itself, the current is not strictly uniformly distributed. It is to be remembered that the ratio of the mean radii of the coils as measured electrically also depends on the cross-section of the coils, and Snow's investigation shows that all the effects of non-uniformity of current-distribution over the cross-section of the winding are eliminated if the ratio of the mean radii is measured electrically. The correction due to the size of the cross-section is, however, by no means negligible in precision work. For example, with the coils originally constructed by Rosa Dorsey and Miller for which the cross-sectional dimensions varied from 1×1 cm. (moving coils) to 2×2 cm., the corrections for the finite size of the cross-section were of the order of 200 parts per million, on the value of the current, while errors of only 0·01 mm. in the determination of the dimensions of the cross-section of the moving coil would cause errors up to 4 parts per million. Since the cross-sectional dimensions of multi-layer coils wound with insulated wire are by their very nature ill-defined, this source of error is of outstanding importance.

Thus having completed measurements with the original coils, Curtis, Curtis and Critchfield[43] made new moving coils for the balance of the National Bureau of Standards, the new coils being specially designed to have accurately defined cross-sections. One of them consisted of a short single-layer solenoid wound with bare wire on a cylinder of pyrex glass. The second coil was wound with anodized aluminium ribbon, that is to say thin aluminium strip, which was coated by electrolytic means with a very thin film of oxide, which provided the necessary insulation. This coil took the form of a flat spiral which was wound under tension on to an aluminium former, the depth of winding being such as to give a square cross-section. Neither of these methods of construction permits the use of as large a number of turns as was provided in the original coils, so that the force to be measured, and therefore the sensitivity of the measurement were somewhat reduced. Nevertheless the accuracy of the final result was considerably improved by the use of the new coils.

Comparisons of the results of absolute measurements are usually somewhat uncertain owing to the lack of precise information about the particular standard cells, resistors and voltameters, in terms of which the various results are expressed. In recent years comparative measurements by the International Bureau of Weights and Measures have established mean international units (p. 307) and it has become possible to compare results by expressing them in terms of the mean international unit as established in 1935. Some comparisons of this kind have been published[44]. They are reproduced in the table below, which also includes the results of the other determinations that have been described, although it must be understood that the comparisons are the more uncertain the greater the time interval between the determination in question and the year 1935.

It will be clear from this table that the uncertainties in our knowledge of the principal electrical units are steadily being reduced, and there is little doubt that when the final results of measurements now in hand in America, Germany, Japan and the U.S.S.R. are available it will be possible to realize the units of the absolute system with an uncertainty of the order of 1 part in 10^5. We have seen that measure-

Results of recent absolute measurements. Values of the quantity, (international unit − absolute unit) expressed in parts per 10^5

Laboratory	Observer	Date	Method	Ohm	Ampere	Volt
N.P.L.	Smith	1914	Lorenz	52	—	—
	Hartshorn and Astbury	1936	Campbell	50	—	36
	Vigoureux	1936	Lorenz	50	—	
	Vigoureux	1935	Ayrton-Jones balance	—	−14	
	Ayrton, Mather and Smith	1908	,,	—	−12	—
N.B.S.	Curtis, Moon and Sparks	1935	A.C. and Maxwell Bridges	46	—	35
	Curtis, Curtis and Critchfield	1936	Rayleigh balance	—	−11	
	Curtis and Curtis	1934	,,	—	− 7	—
	Rosa Dorsey and Miller	1911	,,	—	− 7	—
P.T.R.	Grüneisen and Giebe	1914	A.C. and Maxwell Bridges	48	—	—
L.C.E.	Jouaust, Picard and Héron	1937	Picard	52	—	—

ments to this order of accuracy are apt to become very complicated and laborious in their details, but they have the advantage that when the apparatus has been constructed, the observations can usually be repeated with a precision of a few parts per million. On the other hand when experiments are made with the silver voltameter, the results of successive measurements usually differ by 2 or 3 parts in 10^4, as may be seen by the results of experiments made by H. v. Steinwehr and H. Schulze[45], G. W. Vinal[46], and P. Vigoureux[47] in 1931. It is true that results obtained by taking the mean of many observations agree to perhaps 2 parts per 10^5, but data which depend for their accuracy on such a procedure cannot be regarded as forming a satisfactory basis for modern electrical measurements, and a few physicists will lament the passing of the international units.

REFERENCES

(1) *P. V. Com. int. Poids Mes.* **18**, 130 (1937).
(2) *British Association Reports on Electrical Standards*, p. 60 (Camb. Univ. Press, 1913).
(3) Reference (2), p. 61.
(4) Reference (2), p. 103.
(5) Reference (2), p. 104.
(6) GLAZEBROOK. *Dictionary of Applied Physics*, **2**, 242 (Macmillan, London 1922).
(7) Reference (1), p. 193.
(8) Reference (1), p. 200.
(9) CAMPBELL, N. R. *Proc. Phys. Soc.* **48**, 708 (1936).
(10) RAYLEIGH. *Proc. Roy. Soc.* A, **144**, 266 (1934).
(11) CURTIS, H. L., MOON, C. and SPARKS, C. M. *Bur. Stand. J. Res., Wash.*, **16**, 1 (1936).
(12) JOUAUST, R., PICARD, M. and HÉRON, R. *Bull. Soc. franç. Élect.* **8** (1938).
(13) CURTIS, H. L. *Electrical Measurements*, p. 165 (McGraw Hill, New York and London, 1937).
(14) SNOW, C. *Bur. Stand. J. Res., Wash.*, **7**, 331 (1934).
(15) SNOW, C. *Bur. Stand. J. Res., Wash.*, **9**, 419 (1932).
(16) SNOW, C. *Bur. Stand. J. Res., Wash.*, **6**, 777 (1931).
(17) ASTBURY, N. F. *Proc. Phys. Soc.* **47**, 86 (1935).

(18) ASTBURY, N. F. *Phil. Mag.* **25**, 290 (1938).
(19) SMITH, F. E. *Philos. Trans.* **214**, 27 (1914).
(20) VIGOUREUX, P. *Coll. Res. Nat. Phys. Lab., Lond.* (1938).
(21) GRÜNEISEN and GIEBE. *Ann. Phys., Lpz.*, **368**, 179 (1920).
(22) CAMPBELL, A. *Proc. Roy. Soc.* **87**, 391 (1912).
(23) GIEBE, E. and ZICKNER, G. *Z. InstrumKde*, **53**, 1 (1933).
(24) GIEBE, E. *Congrès International d'Électricité, Paris*, 2nd Section, Paper No. 3 (1932).
(25) CURTIS, H. L., MOON, C. and SPARKS, C. M. *Bur. Stand. J. Res. Wash.* **21**, 375 (1938).
(26) HARTSHORN, L. and ASTBURY, N. F. *Philos. Trans.* **236**, 423 (1937).
(27) CAMPBELL, A. *Proc. Roy. Soc.* A, **107**, 310 (1925).
(28) PICARD, M. *C.R. Acad. Sci., Paris*, **189**, 125 (1925).
(29) NETTLETON, H. R. and LLEWELLYN, F. H. *Proc. Phys. Soc.* **42**, 501 (1930).
(30) NETTLETON, H. R. and LLEWELLYN, F. H. *Proc. Phys. Soc.* **44**, 195 (1932).
(31) NETTLETON, H. R. and BALLS, E. G. *Proc. Phys. Soc.* **47**, 54 (1935).
(32) NETTLETON, H. R. and BALLS, E. G. *Proc. Phys. Soc.* **45**, 545 (1933).
(33) CAMPBELL, A. *Proc. Phys. Soc.* **45**, 553 (1933).
(34) AYRTON, W. E., MATHER, T. and SMITH, F. E. *Philos. Trans.* A, **207**, 463 (1908).
(35) SMITH, F. E. *Dictionary Applied Physics*, **2**, 241 (Macmillan, London, 1922).
(36) VIGOUREUX, P. *Philos. Trans.* A, **236**, 133 (1936).
(37) VIGOUREUX, P. *Coll. Res. Nat. Phys. Lab., Lond.* (1938).
(38) MCCOLLUM, B. *Bull. U.S. Bur. Stand.* (1910).
(39) SMITH, F. E. *Dictionary Applied Physics*, **2**, 239 (Macmillan, London, 1922).
(40) ROSA DORSEY and MILLER. *Bull. U.S. Bur. Stand.* **8**, 269 (1911).
(41) CURTIS, H. L. and CURTIS, R. W. *Bur. Stand. J. Res., Wash.*, **12**, 665 (1934).
(42) SNOW, C. *Bur. Stand. J. Res., Wash.*, **11**, 681 (1933).
(43) CURTIS, H. L., CURTIS, R. W. and CRITCHFIELD, C. L. *P. V. Com. int. Poids Mes.* **18**, 149 (1937).
(44) *P.V. Com. int. Poids Mes.* **18**, 145 (1937).
(45) v. STEINWEHR, H. and SCHULZE, H. *Z. InstrumKde*, **6**, 249 (1932).
(46) VINAL, G. W. *Bur. Stand. J. Res., Wash.*, **8**, 729 (1932).
(47) VIGOUREUX, P. *Coll. Res. Nat. Phys. Lab., Lond.*, **24**, 81 (1932).

ELECTRIC WAVE FILTERS

By N. F. ASTBURY, M.A., F.Inst.P.

§ 1. INTRODUCTION

"The filter stands beside the vacuum tube as one of the two devices making carrier telegraphy and telephony practicable." This quotation, taken from an editorial note on a classical paper by G. A. Campbell[1], expresses the viewpoint of the communication engineer. To the present-day applied physicist these devices are no less important in their incidental applications. The electric wave filter, described briefly by its discoverer as a "device for separating sinusoidal currents of different frequencies", was evolved by G. A. Campbell as a result of the study of the theory of loaded transmission lines. He showed that a line loaded by the insertion of inductive coils gave an improved performance for currents of speech frequencies, while its transmission of currents at higher frequencies was impaired. This important result was followed up by Campbell, and later by O. J. Zobel[2,3] and more recently by Bode[4,5] and by Cauer[6], to quote but a few of the more important contributors to this extensive and growing subject. Amongst treatises on the subject may be noted T. E. Shea's *Transmission Networks and Wave Filters*, K. S. Johnson's *Transmission Circuits for Telephonic Communication*, both published by the D. van Nostrand Co., New York, and A. C. Bartlett's *The Theory of Electric Artificial Lines and Filters* (London, Chapman & Hall). Useful chapters will be found in W. L. Everitt's *Communication Engineering* (New York, McGraw Hill). For the most part these books present the subject on the classical lines of Campbell and Zobel, and this example will be followed here, for although the more recent work presents the subject with a unity of outlook more satisfying to the mathematician, the earlier work is perhaps more readily appreciated by the physicist.

§ 2. OUTLINE OF FILTER THEORY

The description of a filter may now be expanded somewhat. A filter circuit is a network of passive elements which will discriminate between alternating currents of different frequencies, allowing free transmission of currents lying in a chosen frequency band and rejecting currents of frequencies outside this band. When the lower limit of the transmitted frequencies is zero, the filter is described as low-pass and is said to have a cut-off at the upper frequency limit. When the upper limit of the transmitted frequencies is infinite, the filter is described as high-pass and has a cut-off at the lower frequency limit. When both limits are finite, the filter is described as a band-pass. In order to conform to this description, it is necessary that the circuit elements shall be pure reactances, and although in practice these are unobtainable, much of the essential theory has to be worked out on the assumption of no ohmic losses in the circuit. For most practical cases, however, the consequences of ignoring ohmic losses in design calculations are not very serious.

It will, of course, occur to the reader that the familiar resonant combinations of capacitance and inductance exhibit some of the properties characteristic of filters, but they differ from filter circuits in that the latter provide a substantially constant transmission in the accepted band of frequencies, while the simple resonant circuits provide more or less sharp peaks of transmission and rejection. They behave, in fact as band-pass or band-rejector circuits of very narrow band-width.

We may begin the outline of filter theory by considering an infinite transmission line formed by connecting in series a number of identical four-terminal units, built up of symmetrical series and shunt arrangements of circuit elements. Such a line will have a finite input impedance expressing the ratio of applied voltage to current entering the line. If a section n miles long be cut off the line, the input impedance of the remainder will be unaltered, while the input impedance of the n-mile section can be made equal to that of the infinite line by terminating it by an impedance equal to the input impedance of the infinite line. Thus any finite section of line can be made to behave for certain purposes as an infinite line provided suitable terminal impedances are provided. Now suppose each repeated section of the line is made up of three impedances T- or star-connected as in figure 1. Then it can be shown that if Z_0 represent the impedance of such a star looking in at terminals a and b, with no connexion between c and d, and if Z_s is the impedance from the same points with c and d short-circuited, then the input impedance, Z_k, of a line made up of an infinite number of such stars connected in cascade is

$$Z_k = \sqrt{(Z_0 Z_s)}. \qquad \text{.....(1)}$$

This quantity Z_k is also called the "characteristic impedance", "surge impedance" and "iterative impedance", the last term being probably the most descriptive. Evaluating Z_0 and Z_s for the case of figure 1, we find

$$Z_k = \sqrt{\left\{Z_1 Z_2 \left(1 + \frac{Z_1}{4Z_2}\right)\right\}}. \qquad \text{.....(2)}$$

Consider this line terminated by an impedance Z_k as indicated in figure 1. The current in this impedance, which is now providing the load on the line, bears a definite relation to that entering the line, and if ρ is the ratio of the input current to load current, then we define the propagation constant of the line as

$$P = \log_e \rho, \qquad \text{.....(3)}$$

and it can be shown, by an application of Kirchhoff's laws to the circuit, that

$$P = \tanh^{-1} \sqrt{(Z_s/Z_0)} \qquad \text{.....(4)}$$

$$= 2 \sinh^{-1} \tfrac{1}{2} \sqrt{(Z_1/Z_2)}. \qquad \text{.....(5)}$$

If the sections of the line are in π or delta formation, figure 2, results 1, 4 and 5 still hold. The expression for the iterative impedance, Z_k', may be shown to be

$$Z_k' = \sqrt{\{Z_1' Z_2' / (1 + Z_1'/4Z_2')\}}. \qquad \text{.....(6)}$$

It is perhaps worth noting at this stage that mutually equivalent star and delta

combinations can always be found by the use of Kennelly's transformation[7]. The star and delta shown in figure 3 are equivalent if

$$\left.\begin{aligned} X,\ Y,\ Z &= \frac{xy+yz+zx}{x,\ y,\ z}, \\ x,\ y,\ z &= \frac{YZ,\ ZX,\ XY}{X+Y+Z}. \end{aligned}\right\} \qquad \text{.....(7)}$$

In considering the applications of the above results to filter circuits, it must be remembered that the components are taken to be pure reactances. If the reactances are of one sign (i.e. both capacitative or both inductive), Z_k will be a pure reactance, while if they are of opposite signs, Z_k will be a pure resistance. In the latter case, a

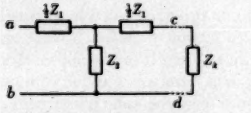


Figure 1. *T* structure.

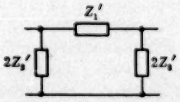


Figure 2. π structure.

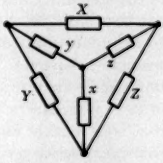


Figure 3. Star-delta transformation.

line correctly terminated with the appropriate Z_k will absorb power from a generator connected at its input end and deliver it in its entirety to the load, since no power can be absorbed in the components. It is easy to see from this that input and output voltages and currents must be equal. If, on the other hand, Z_k is a pure reactance, the line correctly terminated will absorb no power from the generator, since current and voltage will be in quadrature at all points, although attenuation of voltage or current may occur. Further reference to this will be made when the question of losses through incorrect termination is considered.

The nature of the propagation constant must now be investigated. In general, P will be complex and we may put

$$P = 2\sinh^{-1}\tfrac{1}{2}\sqrt{(Z_1/Z_2)} = \alpha + j\beta. \qquad \text{.....(8)}$$

The *attenuation constant* α determines the loss in amplitude of the current, and the *phase constant* β the relative phase of input and output currents. If $Z_1/Z_2 > 0$ (i.e. the reactances are of the same sign) P will be wholly real, α being finite and β zero. Attenuation will thus always occur. This case actually is developed in the theory and design of attenuators for voltage reduction.

If $Z_1/Z_2 < 0$ (i.e. the reactances are of opposite sign) we may put

$$\sinh(\alpha/2 + j\beta/2) = \sinh\frac{\alpha}{2}\cos\frac{\beta}{2} + j\cosh\frac{\alpha}{2}\sin\frac{\beta}{2} = j\tfrac{1}{2}\sqrt{(Z_1 Z_2)}.$$

Thus either $\sinh\alpha/2 = 0$, leading to

$$\alpha = 0, \quad \beta = 2\sin^{-1}\tfrac{1}{2}\sqrt{(Z_1/Z_2)}, \qquad \text{.....(9)}$$

or, $\cos \beta/2 = 0$, leading to

$$\alpha = 2 \cosh^{-1} \tfrac{1}{2} \sqrt{(Z_1/Z_2)}, \qquad \ldots\ldots(10)$$

$$\beta = (2n-1)\,\pi.$$

Thus P takes the forms

$$P = 2 \cosh^{-1} \tfrac{1}{2} \sqrt{(Z_1/Z_2)} + j\,(2n-1)\,\pi$$

$$= 2 \sinh^{-1} \tfrac{1}{2} \sqrt{(Z_1/Z_2 - 4)} + j\,(2n-1)\,\pi, \qquad \ldots\ldots(11)$$

or

$$P = 0 + 2j \sin^{-1} \tfrac{1}{2} \sqrt{(Z_1/Z_2)} = 0 + 2j \cos^{-1} \tfrac{1}{2} \sqrt{(4 - Z_1/Z_2)}. \qquad \ldots\ldots(12)$$

If then $| Z_1/Z_2 |$ lies between 0 and $+4$, the attenuation constant cannot assume a real value and therefore the structure will be non-attenuating, while if $| Z_1/Z_2 | > 4$, the attenuation constant will have a real value. The ratio Z_1/Z_2 is a function of frequency, since Z_1 and Z_2 are of opposite sign, so that it must follow that such a line will pass without attenuation those currents the frequencies of which are such as to make $0 < | Z_1/Z_2 | < 4$, while currents of which the frequencies are such that $| Z_1/Z_2 | > 4$ will be attenuated. This then, is the basis of filter theory, for we have shown that it is possible for a structure to have the property of being non-attenuating for one band of frequencies, while attenuating currents in another band.

It follows that the frequencies bounding the transmission band may be found by solving the equations

$$Z_1/Z_2 = 0, \qquad \ldots\ldots(13)$$

$$Z_1/Z_2 = -4. \qquad \ldots\ldots(14)$$

When Z_1 and Z_2 are each a single inductance or single capacitance, it follows that the solution of the first of these equations is $\omega = 0$ or $\omega = \infty$, $\omega/2\pi$ being the frequency of the alternating current supplied to the line. The solution of the second equation yields a finite value for ω, so that by the use of single inductance and single capacitance elements, it appears possible to build low-pass and high-pass filters as already described. Such filter sections are illustrated in figures 4 (*a*) and 4 (*b*). For figure 4 (*a*), we have $Z_1/Z_2 = \omega^2 L_1 C_1$, so that the transmission band lies between zero frequency and a cut-off frequency given by $\omega = 2/\sqrt{(LC)}$. For figure 4 (*b*), $Z_1/Z_2 = 1/\omega^2 L_2 C_2$, so that the transmission band lies between infinite frequency and a cut-off frequency given by $\omega = 1/2\sqrt{(LC)}$.

When simple low-pass and high-pass filters of equal iterative impedance are joined in series, a band-pass filter is formed, the transmission band of which is bounded by the cut-off frequencies of the two simple filters. By an application of the transformations of equation (7), this combination can be shown to be equivalent to a simple star structure the elements of which are combinations of inductance and capacitance. In this case solutions of equation (13) can be found which are neither 0 nor ∞, as is shown by applying the results to the case of figure 4 (*c*), which represents a simple band-pass filter. In this case

$$Z_1/Z_2 = (1 - \omega^2 L_3 C_3)\, C_4/C_3. \qquad \ldots\ldots(15)$$

The lower bounding frequency is given by

$$\omega_l = 1/\sqrt{(L_3 C_3)}, \qquad \ldots\ldots(16)$$

and the upper by

$$\omega_u = \sqrt{\{(4C_3 + C_4)/C_3 C_4 L_3\}}. \qquad \ldots\ldots(17)$$

In the case of the band-pass filter, a third frequency ω_m, called the mid-frequency, is of importance, for the iterative impedance at this frequency is that used in design; ω_m is the geometric mean of the upper and lower cut-off frequencies. The expression for iterative impedance is naturally more complicated than for the

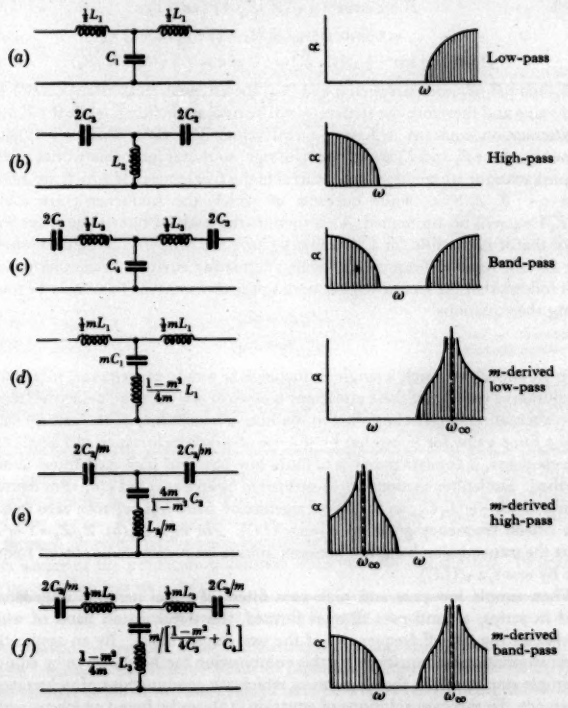


Figure 4. Simple low-, high- and band-pass prototypes and derived forms.

low-pass and high-pass filters, although it is simply evaluated in any given case by applying equation (2).

The attenuation in the attenuating band of the simple filters discussed above is finite, so that the complete rejection of the unwanted frequencies is not possible. If some filter could be found for which, even at one frequency, the attenuation were infinite, then it would be possible, by combining a number of such filters, arranged

to have infinite attenuation at different frequencies, to obtain a filter rejecting completely a whole band of frequencies. Now Zobel[2] showed that it is possible simply to derive from symmetrical "constant-k" networks of the type already considered more complex networks of the same impedance and cut-off but of different propagation constant, and exhibiting infinite impedance at one frequency. These are the so-called m-derived filters, and the idea may be illustrated by the following example. Starting from figure 1, which would be called the prototype, imagine a net of similar structure in which the impedances are Z_1' and Z_2' and let $Z_1' = mZ_1$. Let us now equate the iterative impedances of the derived filter and its prototype. Applying equation (2) we get

$$Z_1 Z_2 \left(1 + \frac{Z_1}{4Z_2}\right) = Z_1' Z_2' \left(1 + \frac{Z_1'}{4Z_2'}\right)$$

$$= mZ_1 Z_2' \left(1 + \frac{mZ_1}{4Z_2'}\right). \qquad \text{.....(18)}$$

Solving for Z_2', we find

$$Z_2' = [Z_2 + (1 - m^2)\, Z_1/4]/m. \qquad \text{.....(19)}$$

Thus, if the derived section has impedances mZ_1 and Z_2' given by (19), it will possess the same impedance and cut-off as its prototype and the structures may therefore be joined together without altering the impedance of the line as a whole, and, as we shall see later, without introducing any reflection losses. It is interesting to note that if the above analysis had been started with the assumption that $Z_2' = mZ_2$, no physically possible value of Z_1 would have resulted which would apply for all frequencies. As it is, m must be less than unity since that portion of the derived shunt impedance represented by $(1 - m^2)\, Z_1/4$ must be of the same sign as Z_1.

Our derived filter has thus a shunt arm of two reactances of different signs. This forms a series resonant path which can be tuned to a frequency near the cut-off frequency. The attenuation of the filter at this frequency then becomes infinite and if this frequency of infinite attenuations is near the cut-off frequency, the cut-off is very much sharper. It is easy to show that if ω_∞ be the resonant frequency of the derived shunt arm and ω_c the cut-off frequency, then

$$\left.\begin{aligned} m^2 &= 1 - (\omega_c/\omega_\infty)^2 \text{ for a low-pass filter,} \\ m^2 &= 1 - (\omega_\infty/\omega_c)^2 \text{ for a high-pass filter.} \end{aligned}\right\} \qquad \text{.....(20)}$$

The derivation of a band-pass filter can be carried out on similar lines, and figures 4*d–f* illustrate m-derived filters of which the prototypes are shown in figures 4*a–c*.

There is another two-element structure used in filter design which is in effect a four-arm bridge. This type, also described as a lattice, is in some ways simpler than the τ and π sections. It is shown in figure 5. The iterative impedance, Z_k, and propagation constant, P, are given by

$$Z_k = \sqrt{(Z_1 Z_2)}, \qquad \text{.....(21)}$$

$$P = \tanh^{-1} 2\sqrt{(Z_1 Z_2)}/(Z_1 + Z_2). \qquad \text{.....(22)}$$

It can be shown that, corresponding to any τ or π section, an equivalent lattice can be found[3]: the converse is not generally true. The lattice type, however, requires more elements than the τ or π types for equivalent results, and on this ground is not so economical. Other features, such as sharpness of cut-off and certain interesting transformation properties, have nevertheless made this type very important in carrier telephony, as will be seen when the subject of crystal filters arises.

The discussion so far has been limited to symmetrical structures. In an unsymmetrical structure, such as shown in figure 6, the input impedance will depend on which end of the line we are looking into. There is thus no unique iterative

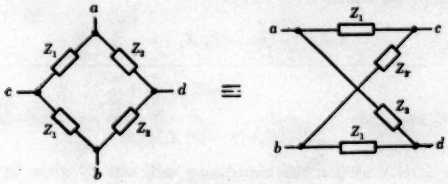


Figure 5. Bridge or lattice structure.

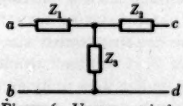


Figure 6. Unsymmetrical τ.

impedance, but there are two analogous constants, the image impedances, such that when a structure is terminated by the appropriate image impedance, the impedances in both directions are identical. If, in figure 6, a, b, be regarded as the input terminals, and if Z_0' and Z_s' are respectively the impedances at these points with cd open- and short-circuited, then the image impedance Z_{ab} for terminals ab will be

$$Z_{ab}=\sqrt{(Z_0'Z_s')}, \qquad \ldots\ldots(23)$$

and if Z_0'', Z_s'' have corresponding meanings for terminals cd, the image impedance for these terminals will be

$$Z_{cd}=\sqrt{(Z_0''Z_s'')}. \qquad \ldots\ldots(24)$$

Instead of the propagation constant, an analogous quantity called the transfer constant is introduced. This is the logarithm of the square root of the ratio of input to output volt-amperes when the structure is appropriately terminated with its image impedances. Calling this T, we have

$$T=\tanh^{-1}\sqrt{(Z_s'/Z_0')}=\tanh^{-1}\sqrt{(Z_s''/Z_0'')}. \qquad \ldots\ldots(25)$$

The similarity between equations 23–25 and equations 1 and 4 should be noted. Detailed development of any particular case will follow on lines similar to those already sketched out for the symmetrical structures.

So far, losses have been ignored, and a complete theory of filters would have to account for losses arising through ohmic dissipation and through incorrect terminations. The former may be dealt with by the introduction of appropriate terms in the expanded forms for Z_1, Z_2, etc., but the resulting expressions are generally cumbersome and not very informative. Both attenuation and phase characteristics are affected by the losses, some of the consequences being obvious from first principles. The attenuation in the pass-bands will attain some small finite value instead of being zero, and in derived filters the attenuation at ω_∞ will never reach infinity. On the whole, the performance in the attenuation bands is less affected by ohmic losses than the performance in the transmission bands.

The losses arising from incorrect termination—the so-called reflection losses—form a branch of transmission line theory into which it is scarcely relevant to enter here, and reference should be made to the books already mentioned for a full treatment. It will suffice to say that when the load into which a transmission unit is working is not identical with the characteristic impedance, the energy arriving at the load will not be absorbed and a reflected wave will travel back to the generator. The chief practical consequence of this in filter circuits is that the attenuation in the pass-bands will become finite.

A neat mechanical demonstration of the behaviour of filters has been given by Lane[8], who constructed a mechanical band-pass filter consisting of a series of pendulums linked together in their common plane of vibration by arched springs. In an interesting series of photographs the attenuation, phase shifting and reflection effects are well illustrated.

It has not been considered worth while here to enter into any detail on three-element filters or filters using mutual inductance, partly because by the application of suitable transformations they can generally be related to τ and π structures and partly because they are more conveniently dealt with as they occur as practical illustrations in the literature.

§ 3. SOME PRACTICAL FILTERS AND THEIR APPLICATIONS

(*a*) *General.* Perhaps one of the most general applications of the filter circuit outside the realm of pure communications has been in combating the effects of the non-linearity of thermionic valve systems. The generation or undistorted transmission of a sinusoidal oscillation by a thermionic valve circuit is, broadly speaking, an impossibility, and the suppression of unwanted harmonics is a problem of some significance. A suitably designed filter circuit—and suitability of design here is largely a question of correct impedance matching—may precede or follow the valve circuit as a whole. For example, in a recent high precision alternating current measurement of the absolute unit of resistance[9], the purity of the supply current was a matter of prime importance, and effective suppression of the harmonics of a 100-c./sec. supply after amplification by a valve was carried out by the use of a composite filter consisting of a τ section, figure 4*a*, followed by a derived section, figure 4*d*. This provides a very straightforward example, and more subtle attacks

on the problem are to be seen in the design of amplifying circuits in which filter units are used for coupling the various stages and for feeding in and out. A useful account of the theory of filter-coupled amplifiers has been given by Butterworth[10]. His paper gives tables for facilitating design of intervalve stages of the general two-element Campbell type. An interesting application of filters in valve circuits is to be found in the coupling stages of beat-frequency oscillators in which a note of audio-frequency is produced as the difference tone between two radio-frequency oscillations. It is customary to connect the mixing valve to the following low-frequency amplifying stages through a low-pass filter, preferably of the composite derived type having a cut-off just above the upper limit of the audio-frequencies generated. It would be desirable to extend the filtering in this case by coupling the radio-oscillators to the mixing circuit through band-pass filters of band-width slightly greater than the range of audio-frequencies generated. As far as the writer is aware, this has not been tried, but it would eliminate the production of audio-frequency harmonics as difference tones of the radio-frequency harmonics and would assist in suppressing hum in cases where the power is derived from the mains.

The suppression of mains hum incidentally provides another example of the use of filter circuits. As a rule, very little design is necessary with filters for this purpose, and only approximate impedance matching need be attempted. It may not be well known, however, that a quite good suppressor unit may be rendered ineffective by improper connexion. It will be obvious that the unit must be a low-pass filter with cut-off below the mains frequency. Terman and Pickles[11] showed that, using a filter unit of this type with a double diode rectifier, the earth-capacitance of the series inductances may shunt the whole filter, forming in fact a bridged-T network of the wrong kind. The remedy is to have no series impedance in the earthed (negative) line, placing the chokes in the positive lead.

The band-pass filter has achieved prominence in recent years by reason of the important part it must play in the efficient operation of radio receivers of the supersonic heterodyne type. In such receivers the band-pass filter is designed to pass the intermediate frequency on to the intermediate-frequency amplifying stages without unwanted oscillations. In this connexion, and in the development of selective radio receivers generally, the use of coupled oscillatory circuits as a band-pass filter has attracted a great deal of attention. If two identical circuits each consisting of an inductance L and capacitance C in parallel are coupled by the mutual inductance M, between the coils, then they will behave as a band-pass filter of which the bounding frequencies of the pass-band are given by

$$\begin{aligned}\omega &= 1/\sqrt{\{C\,(L \pm M)\}} \\ &= \omega_0/\sqrt{(1 \pm k)}, \qquad \text{.....(26)}\end{aligned}$$

where ω_0 is the resonant frequency of either circuit and k is the coupling coefficient ($kL = M$). A discussion of this simple type of filter and its limitations has been given by Howe[12]. One of its limitations is that the band width varies with ω_0, and Bligh[13] has shown how to make the band width approximately constant by making the coupling partly capacitative by the inclusion of a capacitor common to the two

circuits and a capacitor across the open end of the mutual inductance. Theoretical discussions of the coupled circuit type of filter have been given by Oatley[14], by Buffery[15], by Mezey[16] and by Drouin[17].

The construction of a band-pass filter by the connexion of appropriate structures in series has already been mentioned as a theoretical possibility. The double filters of Matsumae and Matsumoto[18] are made by combining filters in parallel, the arrangement being to link up a τ and a π structure. A low-pass filter of this type is shown in figure 7. The filter elements are derived from common values of inductance and capacitance, for if p be the ratio of the characteristic impedance, R_τ, of the τ structure to that, R_π, of the π structure, then

$$\left.\begin{aligned} L_1 &= L_k\,(1+p)/p, & L_2 &= L_k\,(1+p)/2, \\ C_1 &= C_3 = pC_k/2\,(p+1), & C_2 &= C_k\,(p+1), \end{aligned}\right\} \qquad \text{.....(27)}$$

where $L_k = 2R_0/\omega_0$ and $C_k = 2/R_0\omega_0$, ω_0 being the cut-off frequency (which is the

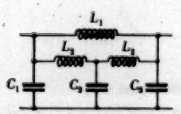


Figure 7. "Double filter."

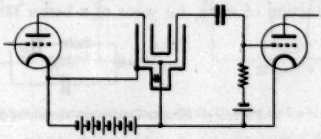


Figure 8. Coaxial cable intervalve filter (Bell Telephone Co.).

same for each component structure) and R_0 being the characteristic impedance given by $1/R_0 = 1/R_\tau + 1/R_\pi$. The virtue of this double filter is that the impedance throughout the transmission band is independent of frequency. Other workers who have proceeded with the double-filter idea are Selach and Zimbalisty[19], whose F-section filters achieve results similar to those described above but use fewer elements.

The use of mutual inductance in filters of very narrow band-width is illustrated by a whistle filter described by P. K. Turner[20]. This, in effect, is an elaboration of Albert Campbell's well-known mutual-inductance-capacity sifter circuit[21] for cutting out one frequency. Turner's circuit is an ingenious modification of the circuit to an intervalve stage, additional components being introduced to produce what is in fact a two-phase resonance bridge balance for suppression of the unwanted note. Resonance bridges of various kinds can, and in fact have been, used as filters for one frequency, particularly in harmonic analysis circuits[22].

In designing filters for very high frequencies the difficulties of providing efficient chokes and condensers of conventional design become very great. Mason and Sykes[23], of the Bell Telephone Laboratories, have shown that coaxial cable may be adapted for filter units, the ratio of reactance to resistance in such a cable being very high. One filter described for valve-coupling and shown in figure 8 consists of a length $2l$ of cable shunted at its mid-point with a length l in which the middle and outer conductors are joined together at the remote end. This is really a multiple band-pass unit in its simple form, but by adding condensers single band-pass, low-

pass and high-pass filters are obtainable. The cable filters described are used for frequencies ranging from 50 to 150 Mc./sec. The high quality of the filters is shown by the fact that one model had a loss at mid-band of only 1 decibel, while the loss 2 Mc. on either side had increased to 50 db.

Before leaving the purely electrical type of filter, it is interesting to note that by departing from the condition that a filter shall consist only of passive elements, Chakravarti[24] has shown that an amplifying filter, giving a voltage gain in the transmission band, may be produced. It appears from theoretical considerations that, if a filter be terminated by a *negative* resistance of magnitude equal to the iterative impedance, then this effect is produced. A suitable negative resistance is provided by a screened grid valve operating under the condition of secondary emission. Control over the anode impedance in this condition is limited and the circuit is arranged for adjustment by the addition of a resistance in series with the valve.

(*b*) *Electromechanical filters.* The most recent development in filter circuits is the production of what, for want of a better term, may be described as electro-

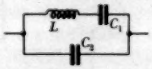


Figure 9. Simple equivalent circuit of piezo-electric crystal.

mechanical filters. Like the original filters, these too were developed in the Bell Telephone Laboratories, W. P. Mason[25] being largely responsible. With all purely electrical circuits the sharpness of cut-off is limited by the losses in the reactances. A possible escape from this difficulty lies in the use of a mechanical element instead of an electrical inductance, so coupled to the electric circuit that its inertia is made to act as an inductance. A piezo-electric crystal such, for example, as quartz, provides just such an element. The simple equivalent circuit of a quartz plate between plane electrodes is shown in figure 9. Its inertia provides the inductance, its stiffness (elasticity) a capacitance in series with this inductance, while the inter-electrode capacitance shunts the whole. The damping, which would represent resistance losses, will be extremely small. Such a crystal, or four crystals, say, in a lattice structure, would give a very narrow pass-band limited in width to rather less than 1 per cent of the transmitted frequency. Mason showed that the addition of an inductance of the ordinary kind in series with each crystal in the lattice would give up to 15-fold broadening of the band. Although at first sight the introduction of such inductances would appear to defeat the whole object of the crystal filter, it can be shown, and for a full discussion Mason's paper should be consulted, that in a lattice network of the type described, the resistance of the coils is brought in effect to the ends of the lattice and incorporated with the terminal resistances, so that the effect of the dissipation in the coils is to add a constant loss, independent of frequency, to the filter characteristic. In other words, transmitted and attenuated bands are equally affected. Similar effects can be obtained by the use of shunt inductances, a scheme particularly useful for intervalve coupling filters where high iterative impedances may be desirable.

A quartz plate, as is well known, possesses several modes of vibration—two longitudinal parallel to its edges, one perpendicular to its surface, one flexural and one shear or torsional—and it is necessary that these modes shall have well differentiated frequencies so that coupling between them is avoided. Mason's first experiments were made with crystals cut with their thickness parallel to the electric axis, their width parallel to the optic axis and their length parallel to the mechanical axis. In addition to the desired longitudinal mode, such crystals showed a strong tendency to produce a torsional mode at a 40 per cent higher frequency, giving in consequence an unwanted pass-band. From an analysis of the mechanics of the

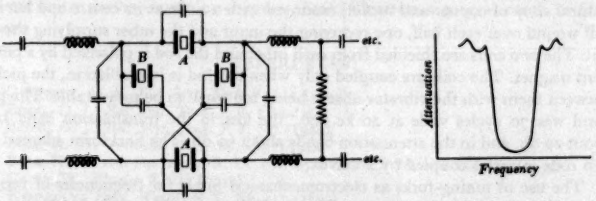


Figure 10. Lattice crystal filter for carrier telephony (Bell Telephone Co.).

problem Mason showed that this could be avoided by cutting the plate with its face perpendicular to the electric axis, but with its ends and sides making a small angle with the mechanical and optical axes. The major axis of the plate is in fact inclined at an angle of 18·5° to the mechanical axis of the crystal. A crystal filter unit of the type now in service in multi-channel telephony is shown, with its attenuation curve, in figure 10[26]. Although four crystals are shown in the diagram, there are only two in reality, those marked *AA* and *BB* being actually provided by one crystal with bifurcated electrodes. The crystals are from 2·4 to 4·2 cm. long, about half this in width and one-fiftieth of this in thickness. The electrodes are made by coating the crystal with aluminium by vacuum evaporation. The coils have molybdenum permalloy dust cores.

The crystal filter is also used in the more advanced types of superheterodyne receiver. The first detector valve is usually transformer-coupled to a closed circuit of which a single crystal forms a series element. This circuit is capacitatively coupled to an auto-transformer forming part of the tuned grid circuit of the first intermediate-frequency amplifying valve. Selectivity and band-width control, which is of course limited, are provided by series and shunt capacitances in the coupling circuit. The very high selectivity of this arrangement is admirable for telegraphy reception, but for telephony the sharpness may be so fine as to cause undesirable attenuation in the speech side-bands, leading to a loss of quality and even of intelligibility. A somewhat less selective filter for telephony which is, however, still much superior to the purely electrical circuit, is provided by the Brush Transfilter, which is a composite affair in which a short steel rod, $\frac{3}{4}$ in. long and $\frac{5}{16}$ in. in diameter, is driven by two tiny plates of Rochelle salt, 0·125 in. sq. and 0·02 in. thick. Two similar plates act

as pick-ups. The arrangement is such that the steel bar more or less takes charge of the assembly, any variations in crystal characteristics having little effect on the performance of the complete unit. For a description of various types of crystal filter used in radio receivers the reader is referred to *The Radio Amateur's Handbook*, published by the American Radio Relay League.

The difficulties of producing simple quartz plates for frequencies less than, say, 50 kc./sec. limits the use of crystal filters to higher frequencies. However, there is available for lower frequencies an analogous electromechanical system in the magnetostriction oscillator. This has been successfully applied by H. H. Hall [27] to the production of a 20 kc./sec. filter. He used a half-wave rod of monel metal (a natural alloy of copper and nickel) mounted with a node at its centre and having a coil wound over each half, one receiving the input and the other supplying the output. The two coils are shielded from each other and the rod is polarized by a permanent magnet. The coils are coupled only when the rod is in oscillation, the pick-up between them with the vibrator absent being too small to be measurable. The pass-band was 70 cycles wide at 20 kc./sec., the loss in the transmission band being about 20 db. and in the attenuation bands about 60 db. The final form adopted uses two rods in series coupled by a valve.

The use of tuning-forks as electromechanical filters for frequencies of 100 and 1000 c./sec. has been described by Jouaust, Picard and Héron [28]. In their work at the Laboratoire Central d'Électricité on the absolute measurement of resistance, a working supply at a frequency of 100 c./sec. was obtained by demultiplication of a 1000 c./sec. supply transmitted along a telephone line from some distance. This supply was coupled to the demultiplying circuit through a tuning fork of frequency 1000 c./sec., thus eliminating not only harmonics but also parasitic voltages picked up by the line, while the output from the demultiplier was coupled through a tuning fork of frequency 100 c./sec. to a normal electrical filter and thence to amplifiers, etc. The attenuation of the low-frequency filter system was 30 db. at 150 c./sec. and 70 db. at 200 c./sec.

(*c*) *Phase-shifting and constant-current circuits.* In conclusion we may note one or two incidental properties of electrical filter circuits which have interesting special applications. If the elements of a lattice or bridge structure, figure 5, are made reciprocal to a resistance R, then the iterative impedance is equal to R and is independent of frequency, equation (21). The transmission band is thus of infinite width, but phase-shift will be produced between input and output. This then raises the possibility of non-attenuating phase-shifting circuits [29]. If in figure 5 $Z_1 = j\omega L$ and $Z_2 = 1/j\omega c$, then $R = \sqrt{(L/C)}$. When a voltage of constant amplitude V and frequency $\omega/2\pi$ is applied to terminals ab and a resistance R joins cd, then the current in R will be V/R at all frequencies, but differing in phase from V by an angle θ given by

$$\theta = 2 \tan^{-1} \omega \sqrt{(LC)}. \qquad \text{.....(28)}$$

Other more complicated circuits of the same type can be devised having similar properties.

Certain filters possess the property that at some frequencies in the transmission bands they will deliver into *any* real load a current of constant amplitude provided

that the amplitude of the input voltage remains constant. This is the basis of Boucherot's constant-current networks[30]. A simple example derived from the τ structures of figures 4*a* and 4*b* is provided by an inductance L and a capacitance C in series, with a load Z connected in shunt across L or C. A voltage of constant amplitude V and frequency $\omega/2\pi$ applied to the ends of the L-C combination will produce a current in Z of amplitude $V\sqrt{(C/L)}$ when the frequency is such that $\omega^2 LC = 1$, i.e. the current in Z is independent of Z. The phase-shifting bridge just mentioned possesses similar properties. An impedance Z connected between c and d, figure 5, receives a current $V\sqrt{(C/L)}$ when the input voltage is V and its frequency is again given by $\omega^2 LC = 1$. These properties, which are easily derived from first principles in the simple cases given above, illustrate a general result in filter theory.

REFERENCES

(1) Campbell, G. A. *Bell Syst. Tech. J.* **1**, 1 (1922).
(2) Zobel, O. J. *Bell Syst. Tech. J.* **2**, 1 (1923).
(3) Zobel, O. J. *Bell Syst. Tech. J.* **3**, 567 (1924).
(4) Bode, H. W. *J. Math. Phys.* **13**, 275 (1934).
(5) Bode, H. W. and Dietzold, R. L. *Bell Syst. Tech. J.* **14**, 211 and 215 (1935).
(6) Cauer, W. *Physics*, **2**, 243 (1932).
(7) Kennelly. *Elect. World, N.Y.*, **34**, 413 (1899).
(8) Lane. *Elect. Engr*, **52**, 813 (1933).
(9) Hartshorn, L. and Astbury, N. F. *Philos. Trans.* A. No. 769, **236**, 423 (1937).
(10) Butterworth. *Exp. Wireless*, **7**, 536 (1930).
(11) Terman and Pickles. *Proc. Inst. Radio Engrs, N.Y.*, **22**, 1040 (1934).
(12) Howe. *Exp. Wireless*, **8**, 233 (1931); *Wireless Engr*, **14**, 289 and 347 (1937).
(13) Bligh. *Wireless Engr*, **9**, 61 (1932).
(14) Oatley. *Wireless Engr*, **9**, 608 (1932).
(15) Buffery. *Wireless Engr*, **9**, 504 (1932).
(16) Mezey. *Onde élect.* **12**, 149 (1933).
(17) Drouin. *Onde élect.* **11** (May, June, July and August 1932).
(18) Matsumae and Matsumoto. *Elekt. Nachr.-Tech.* **11**, 172 (1934).
(19) Selach and Zimbalisty. *Elekt. Nachr.-Tech.* **12**, 243 (1935).
(20) Turner, P. K. *Wireless World*, p. 588 (Dec. 1936).
(21) Campbell, A. *Proc. Phys. Soc.* **21**, 80 (1908).
(22) Belfils. *Rev. gén. Elect.* **31**, 267 (1932).
(23) Mason and Sykes. *Bell Syst. Tech. J.* **16**, 275 (1937).
(24) Chakravarti. *Proc. Indian Acad. Sci., Bangalore*, **1** A, 224 (1934); *Onde élect.* **15**, 150 (1936); *Phil. Mag.* **26**, 173 (1938).
(25) Mason, W. P. *Bell Syst. Tech. J.* **13**, 405 (1934). (See also Ralph, *Elect. Commun.* **11**, 204 (1932–3)).
(26) Buckley. *J. Appl. Phys.* **8**, 40 (1937).
(27) Hall. *Proc. Inst. Radio Engrs, N.Y.*, **21**, 1328 (1933).
(28) Jouaust, Picard and Héron. *Trav. Lab. cent. d'Élect., Paris* (July 1938).
(29) Bartlett's book, p. 117.
(30) Bartlett's book, p. 120.

The following papers, not specifically referred to in the text, are extremely comprehensive and give numerous circuits and performance data:

(31) Johnson and Shea. *Bell Syst. Tech. J.* **4**, 52 (Jan. 1925).
(32) Little, G. J. S. *Pap. Instn Post Off. Elect. Engrs*, paper No. 143 (1931).
(33) Halsey, R. J. *Pap. Instn Post Off. Elect. Engrs*, paper No. 147 (1932).

A digest of classical filter theory, with illustrative numerical examples, is given by M. G. Reed in a series of articles in *Exp. Wireless*, **7** (1930).

ATOMIC PHYSICS

I. THE SINGLE SCATTERING OF ELEMENTARY PARTICLES BY MATTER

By F. C. CHAMPION, M.A., Ph.D.
University of London, King's College

§1. INTRODUCTION

ALTHOUGH Rutherford's pioneer work on the single scattering of α particles was basically responsible for such an important conception as the nuclear theory of matter, the subject of scattering has in recent years suffered something of an eclipse. This was due to the concentration of effort partly on nuclear disintegration and partly on the elucidation of the nature of new fundamental particles. However, the difficulties experienced by the theoretical physicist in building satisfactory nuclei from neutrons and protons soon showed the imperative need for much more experiment on single scattering to assist in the determination of nuclear forces: for the scattering of particles is one of the best methods of examining nuclear fields. Rutherford and Chadwick, and Chadwick and Bieler showed quite early that for fast α particles incident on light nuclei, the scattering was not in agreement with Rutherford's classical scattering law. In order to account for the deviations from Coulomb's law it became necessary to ascribe a structure to the particles. With the advent of wave mechanics, however, this concept of structure was replaced by a potential hole of fairly definite depth and width, and this potential, when introduced into the wave equation, gave a fair representation of the experimental facts.

Still more recently, Bohr[1] and others have pointed out that it is not legitimate to treat the interaction of particles with nuclei as a one-body problem, that is, as the problem of the scattered particle moving simply in the potential field of the scattering nucleus considered as a whole. The new ideas were evolved from considerations of very close collisions, in fact, from nuclear disintegration, but it is now clear that a general treatment must include both nuclear disintegration and scattering. One of the main features of nuclear disintegration is that as the energy of the incident particles is raised, for certain values there is a large increase in the amount of disintegration, while at still higher energies a continuous increase is observed. Such behaviour is explained on the assumption that for the lower energies resonance levels must exist, the eventual steady increase being attributed to the passage of the incident particle over the top of the nuclear potential barrier. These peculiarities in the disintegration phenomena have their exact counterpart in the single scattering of incident particles, deviations from Rutherford's law being attributed to these resonance levels and to the particular value of the height of the potential barrier. The general formula proposed by Bohr and others to cover both disintegration and scattering phenomena bears a marked resemblance to the dispersion formula for the

scattering of light. The formula also represents correctly the asymptotic behaviour of the wave-functions of a particle in a Coulomb field. This is particularly essential since the contributions of high orbital momenta are very important in Coulomb scattering. Unfortunately the expressions obtained from Bohr's many-body theory are rather complicated but in special cases they take a simpler form. Indeed it is found that the scattering by light nuclei when treated as a one-body problem, assuming a suitable potential to act between the scattering nucleus and the scattered particle, yields results which are not very different from those given by the more general theory. Such agreement is not accidental, for if, for example, the spins of the scattered and scattering particles are zero, the scattered wave contains only partial waves of zero magnetic quantum number. It should be mentioned that caution should be used in interpreting the results obtained from scattering data. Resonance may occur for energies above the barrier, penetration continues to increase at higher energies for particles of higher angular momenta, and the probability of penetration of the barrier is quite large even when the incident particles have energies lower than the actual barrier height.

We shall consider, in order, the scattering of protons, neutrons, deuterons, α-particles and electrons. Neither deuterons nor α-particles are, of course, strictly speaking, elementary particles, but the study of their scattering throws so much light on the nature of nuclear fields that their inclusion is justified. No experimental work has yet been carried out on the single scattering of neutrinos, neutrettos, yukons, barytrons nor heavy electrons. We shall also exclude the scattering of photons from our discussion.

§2. PROTONS

Hafstad, Heydenburg and Tuve[2] have continued their work on the scattering of protons by protons. They have established the existence of large attractive nuclear forces between protons at close distances, in addition to the usual Coulomb repulsion. In the absence of these attractive forces the scattering should be given by the well-known formula due to Mott

$$dQ(\theta) = \frac{\pi e^4}{E_0^2}(\operatorname{cosec}^4\theta + \sec^4\theta - \operatorname{cosec}^2\theta\sec^2\theta)\sin\theta\cos\theta\,d\theta,$$

where $dQ(\theta)$ is the cross-section for a deflexion through an angle between θ and $(\theta+d\theta)$, e is the electronic charge and E_0 is the energy of the incident protons. In figure 1, the ratio of the observed scattering to the Mott scattering is given as a function of the energy of the incident protons for an angle of scattering of 45°.

At low energies, that is below 200 kv., the scattering is in agreement with Mott's formula. As the energy is raised, however, the attractive forces become effective and have almost neutralized the Coulomb repulsion at about 400 kv. The attractive forces then preponderate and the scattering rises to the Mott value again at about 650 kv. The scattering then continues to rise apparently indefinitely as the energy is raised still further. The experimental results have been accounted for theoretically by Breit, Condon and Present[3] and, in a recent paper, Eddington[4] agrees essentially with their conclusions. We may observe at this point that Born[5] has considered

the relations of relativity and quantum theory and has been led to formulate a principle of reciprocity. From this principle he deduces that Rutherford's scattering law must be changed by a factor

$$\left(1-\frac{E}{E_0}\sin^2\frac{\theta}{2}\right)^{-1}$$

where E is the energy of the incident particle and E_0 is a constant energy equal to about 10·4 mv. for proton-proton collisions. The results are stated to be in agreement with the experimental values of Hafstad, Heydenburg and Tuve, and several applications are made to other problems in atomic and nuclear physics.

Considering now the scattering of protons by deuterons, no further experimental work seems to have been done since the earlier work of Tuve, Heydenburg and Hafstad[6]. On the theoretical side Ochiai[7] has attempted to apply Schiff's method[8]

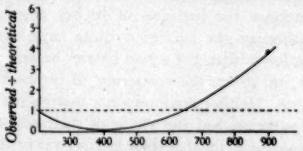


Figure 1.

for neutron-deuteron collisions to proton-deuteron collisions. The theoretical results are not in very good agreement with experiment. However, Primakoff[9] gives reasons for believing that the large excess scattering of protons by deuterium, which is observed experimentally at an energy of about 0·88 mv., is due to resonance scattering of protons with orbital momentum greater than zero. Ochiai does not consider that Primakoff's treatment accounts for all the discrepancy. It may be stated that Primakoff used the one-body theory, justifying his procedure by noting that since capture and inelastic scattering are comparatively rare, the use of the many-body theory is not likely to lead to appreciably different results. It would appear that there is room for considerably more experimental and theoretical work on the scattering of protons by deuterons.

Recently the scattering of protons by helium has been examined experimentally by Heydenburg and Roberts[10] between angles of 20 and 40° for proton energies of 700 and 100 kv. The results are in agreement with Rutherford's scattering law and thus confirm the earlier experiments of Chadwick and Bieler on the scattering of slow α-particles by hydrogen nuclei.

§3. NEUTRONS

The scattering of neutrons, like the scattering of protons, assumes particular importance since neutrons are fundamental constituents of nuclei. They have, however, the important distinction that whereas slow protons are scattered by the Coulomb field at some distance from the nuclear centre, slow neutrons experience

no such repulsive field and their scattering will therefore often result from a close interaction with the scattering nucleus. For the purposes of convenience we shall divide the subject into two groups according as the neutrons are fast or slow.

Considering first the scattering of fast neutrons the position is as follows. Regarding the earlier results, Kurie used an expansion chamber to examine the angular distribution of protons ejected from paraffin wax by neutrons from a source of (Po + Be). In ordinary laboratory coordinates, a predominance in the scattering at 45° indicates isotropic scattering in a system of coordinates moving with the centre of gravity of the neutron-proton system. Kurie found wide departures from isotropy and theoretical work of Morse, Fisk and Schiff showed that small but finite values of the depth of the potential hole which represents the proton would lead to anisotropic scattering. However, for other reasons it was felt that the hole should be of zero depth, and Bethe and Bacher showed that for a two-dimensional hole the scattering should be isotropic until the energy of the neutrons was greater than about 20 mV. The experimental work was continued by Meitner and Philipp who used a neutron source similar to that used by Kurie but who examined the proton tracks projected in hydrogen gas in an expansion chamber. The distribution was isotropic and similar results were obtained by Auger and by Monod-Herzen. On the other hand, a more detailed study by Harkins, Kamen, Newson and Gans, who used neutron sources of Be mixed with mesothorium and thorium X and in some cases with radiothorium, again found the distribution anisotropic. These earlier results have been criticized by Dee and Gilbert[11], partly on the grounds of the inhomogeneity of the velocities of the neutrons used and partly for other reasons. These authors used the monokinetic beam of neutrons obtained by the deuteron-deuteron reaction to examine the scattering of 2·4 mv. neutrons as registered by 2000 proton recoil tracks. Their results showed very satisfactorily that the scattering was isotropic, and a theoretical analysis of the position by Massey and Buckingham[12] supports this conclusion. Confirmatory experimental results using the monokinetic D—D neutrons have also been obtained by Bonner[13] and by Kruger, Shoupp and Stallmann[14], all of whom find spherical symmetry. The former, indeed, points out that spherical symmetry is to be expected since the de Broglie wave-length is large compared with the currently accepted range of the neutron-proton forces.

We may pause to note, as emphasized by Heydenburg and Tuve, that there is found to be close agreement between the empirical values of proton-proton and proton-neutron interactions in ^1S states, suggesting that apart from the different Coulomb and spin interactions, the interaction between fundamental particles is independent of their charge. The apparent preference for equal numbers of neutrons and protons in the construction of nuclei is therefore conditioned more by the operation of Pauli's principle than by the greater value of neutron-proton as compared with proton-proton forces. Indeed, calculations of Breit and Stehn[15] lead to the conclusion that there is very little difference in the magnitude of neutron-proton and proton-proton interactions, although the former may be slightly greater.

With regard to the scattering of slow neutrons by protons, the scattering of

neutrons by ortho- and para-hydrogen has been examined mathematically by Schwinger and Teller[16]. The results provide evidence on the sign of the binding energy of the deuteron singlet state, on the range of the neutron-proton interaction and on the dependence of that interaction on spin orientation. Neutrons with energy less than 0·068 ev., incident upon para-hydrogen in the ground state, may be either elastically scattered or inelastically scattered with excitation to the ground state of the ortho-system. This latter process requires 0·032 ev. energy and occurs only if the neutron-proton interaction depends on spin. When the neutron energy is less than 0·045 ev. the cross-section for the scattering of the neutrons by ortho-hydrogen in its ground state will be the sum of the elastic scattering cross-section and the cross-section for the inelastic process in which the molecule is converted to a para-molecule in its ground state with the neutron taking up the excess energy. The cross-sections of these four processes have been calculated, assuming an interaction range of 2×10^{-13} cm. and a virtual singlet state of the deuteron. For 73° K. neutrons, $\sigma_{para} = 0{\cdot}21 \times 10^{-24}$ cm^2, while $\sigma_{ortho} = 65 \times 10^{-24}$ cm^2. The cross-sections for 293° K. neutrons, however, are 19×10^{-24} cm^2 and 50×10^{-24} cm^2 respectively. Hence, if the present concept of the neutron-proton interaction is valid it is to be expected that at liquid air temperatures the scattering of neutrons by ortho-hydrogen should be about 300 times that by para-hydrogen. Further, the scattering of neutrons by para-hydrogen at room temperature is 100 times its value at liquid air temperatures. Experiments of Halpern, Estermann, Simpson and Stern[17] show that, as predicted by Schwinger and Teller, the neutron scattering by ortho-hydrogen is much greater than that by para-hydrogen at liquid air temperatures. Since the ratio of the cross-sections is approximately unity if the singlet state of the deuteron were real, the virtual nature of this state seems definitely established. Assuming this as proved, Schwinger and Teller show that the cross-section for elastic scattering by para-hydrogen is sensitive to the range of the neutron-proton interaction. For example, the cross-section at 73° K. with zero range of interaction is $1{\cdot}75 \times 10^{-24}$ cm^2, whereas it is $0{\cdot}26 \times 10^{-24}$ cm^2 for a range of 2×10^{-13} cm. It should be observed that intermolecular forces have been ignored in these calculations but their neglect seems to be justified according to the preliminary results of Carroll, Powers, Berger and Dunning[18]. These workers examined the scattering of 300° K. neutrons by the protons in H_2 and CH_4 respectively and found that the effective cross-section changed only from about 35×10^{-24} cm^2 to 45×10^{-24} cm^2.

Turning now to the scattering of neutrons by deuterons we may observe that owing to the similar nature of the neutron and proton, the large excess scattering of 0·88 mv. protons by deuterium should have its counterpart in a large excess scattering for neutrons of this energy. Unfortunately a monokinetic beam of neutrons of this energy has not yet been devised experimentally. However, the angular distribution of 328 recoil deuterons due to the impact of 2·8 mv. neutrons from the D—D reaction has been examined by Kruger, Shoupp, Watson and Stallmann[19] using an expansion chamber filled with 54 per cent D, 37 per cent A, 6 per cent N and 3 per cent heavy water vapour. The results show marked deviation from spherical symmetry, a prominent peak being observed at 60°. The authors do not attempt a

detailed theoretical explanation of this divergence from S scattering. However, they point out that their results could be accounted for in at least three ways. First, fair agreement is obtained with the theoretical curves of Massey and Mohr[20], who take into account also the P scattering. There is some discrepancy at small angles, but the choice of a slightly different potential hole more in accordance with the probable structure of H_3 might give complete agreement. Secondly, polarization of the deuteron by the neutron has not been considered. Finally, it is possible that higher orders of the partial waves must be considered in the description of the scattered neutron wave.

We must next review the evidence for the inelastic scattering of neutrons by nuclei. We refer here to inelastic collisions resulting in the emission of a γ ray. Seaborg, Gibson and Grahame[21] reported what they considered to be such inelastic scattering of neutrons and came to the conclusion that the majority of the low energy γ radiation emitted by elements under neutron bombardment arose in this way. Using the fast neutrons from a Ra–Be source, the average energy of the γ rays produced when the scatterer was lead was just under 1 mv. It was further shown that unless the neutrons had an energy greater than about 0·5 mv., this inelastic scattering process did not occur. On general grounds it does not seem immediately obvious how a neutral particle would produce γ radiation by nuclear impact of any simple type and, in a later paper, Grahame and Seaborg[22] state that the majority of the γ radiation probably arises from momentary amalgamation of the neutron with the nucleus and subsequent emission of a neutron and a nuclear γ ray. They suggest, however, that small energy losses may be due to weak nuclear interaction which they term surface scattering. These results have received indirect support from the experiments of Hudspeth and Bonner[23] on the inelastic scattering of 2·5 mv. neutrons by lead. The energies of the neutrons after scattering were found from the recoil energies of protons in the cloud chamber. The results showed that 10–15 per cent of the neutrons made inelastic collisions, losing from 0·5 to 1·3 mv. energy in the process. On the other hand, the maximum energy of the neutrons was unaffected by the lead to within the experimental error of 0·1 mv.

Other evidence of inelastic scattering has been obtained by Kikuchi, Aoki and Husimi[24], although it is probable that the interpretation which they originally placed on their results is in error. They describe what they term the neutro-electric effect. This consists in the appearance of a proton and an electron track starting from the same point in the gas in an expansion chamber when the latter is traversed by fast neutrons. Initially they suggested that a proton could disintegrate into a proton and an electron, but experiments of Gilbert, Smith and Fremlin[25] have shown this suggestion to be definitely incorrect. They next proposed that the effect was due to the recoil of a nucleus and the simultaneous ejection of an electron from the same atom, the energy for both processes being derived from the energy of the incident electron. To test this suggestion further they examined the tracks produced in a cloud chamber filled with methane and traversed by the 2·6 mv. neutrons from the D—D reaction. The phenomenon was again observed but to a much smaller extent and the possibility of the effect being due to coincidence has to be considered. The

experiments are discussed by Seaborg, Gibson and Grahame who state that it is possible that a few per cent of the ionization observed by them and attributed to γ radiation might be a bona-fide neutro-electric effect. Clearly further experimental and theoretical investigation of this problem are required.

To conclude this discussion on the scattering of neutrons we may note that Goldhaber and Briggs[26] have examined the scattering of slow neutrons by a great many elements and observe that while the scattering cross-section varies by a factor of about 10, the absorption cross-section varies by a factor of 10^6. This general behaviour is in accordance with existing theory in which the phenomena are treated as many-body processes. A suggestion which has been made that the isotopic displacement of spectral lines might be attributed to neutron-electron interaction has been shown by these experiments of Goldhaber and Briggs to be untenable. The cross-section for the scattering of slow neutrons by electrons is only about 10^{-27} cm^2 and the corresponding interaction is far too small to affect the spectral lines.

§ 4. DEUTERONS

While the deuteron cannot claim the distinction of being an elementary particle it has the unique position of being the simplest type of compound nucleus. Its nuclear field is therefore of considerable interest and experiments on the scattering of deuterons by deuterium and by helium have been carried out by Heydenburg and Roberts. With the former, the results show an excess above the Mott scattering in the ratio of 1·3 at 15°, and 2·8, 3·25 and 4·1 at 45° for incident energies of 640, 743 and 855 kv. respectively. This behaviour can be explained on the assumption of an additional repulsive force between deuterons for close distances of approach. This is directly the reverse of the proton-proton forces which, as we have seen, are found to be attractive. For helium, the ratio of the observed scattering to the Rutherford value is unity at 20° rising to twice at 55° for deuterons of 667 kv. energy, and unity at 20° rising to 4·45 at 75° for deuterons of energy 887 kv. The full interpretation of these results is not yet available.

§ 5. α PARTICLES

The various formulae for elastic scattering are, as the name implies, based upon the assumption that energy and momentum are conserved in the scattering process. That this is true in the collisions of α particles with hydrogen nuclei was shown quite early by Blackett and Lees. Further, unless nuclear disintegration occurs, no examples of inelastic collisions of α particles with any type of nucleus have yet been satisfactorily established with the expansion chamber although the emission of some γ-radiation has often been detected electrically when α particles impinge on solid bodies. In this connexion, during the past year, McCarthy[27] has taken 15,000 photographs of the α particles from Th (C + C′) yielding a total of 600,000 tracks in neon and deuterium. Of the many collisions obtained, all were found to be perfectly elastic, not even any examples of nuclear disintegration being observed although a few such cases were to be expected in the amount of data obtained. It may also be observed that the scattering of α particles by the heavy elements Ag, Sn, Sb and Te

has been found to be in agreement with Rutherford's law. These experiments were carried out by Krammer[28] and corresponded to distances of approach of 1·84, 1·97, 2·00 and 2·04 in units of 10^{-12} cm.

For light nuclei, measurements have been made by Brubaker[29] on the scattering of α particles between 35° and 130° by argon, oxygen and neon. A new type of apparatus in which the particles are detected by annular proportional counters has been developed, which makes it possible to study two ranges of angle simultaneously. Anomalies have been observed in the scattering by all three elements. The results of Riezler for oxygen and neon have been confirmed and extended to higher particle energies and larger angles. At the larger angles, the ratio of the observed to the classical scattering for oxygen rises to 4 at 6·7 mv., which is higher than can be explained by S scattering alone. With neon the ratio of the observed to the classical scattering falls below unity as the energy of the incident particles is raised but rises again to unity at 6·9 mv. For argon the anomaly consists in a decrease below the classical value beginning at the highest angles for an energy of 5·5 mv. and continuing to the highest energies used. The form of the anomaly indicates that it should be attributed to overcoming the potential barrier although it occurs at a lower energy than the generally accepted barrier height would lead one to expect.

Experiments on the scattering of particles at low energies have been made by Gerthsen[30] on the scattering of helium canal rays by helium nuclei for a velocity of $3 \cdot 10^8$ cm./sec. for angles from 15° to 45°. The scattering in air was used as a check on the results; the values obtained are shown in table 1. The first column contains the ratio of the angles of scattering, while the second and third columns give the ratio of the scattering to be expected on classical theory according to whether the scattered and scattering particles are different or identical in nature. In column 4 is given the ratio to be expected on wave-mechanical theory for similar particles as calculated from Mott's formula, while columns 5 and 6 contain the observed ratios in helium and in air respectively.

Table 1

θ_1/θ_2	Theory			Observation	
	Classical		W.M. Sim.	He in He	He in air
	Diff.	Sim.			
15/20	4·15	4·08	3·25	2·75	3·81
20/25	2·99	2·91	4·78	3·46	2·87
25/30	2·42	2·31	4·35	3·88	—
30/45	4·78	3·82	0·77	1·18	5·0

Examination of the table shows that the observed results are in best agreement with wave-mechanical theory. Departures are ascribed to heavy impurities or the effect of electron screening and are not considered to represent any real divergence between theory and experiment. The results therefore confirm those obtained some time previously with slow α particles by Blackett and Champion for the same region of velocities.

§ 6. ELECTRONS

We shall confine our discussion mainly to the single scattering of electrons with energies of more than 0·1 mv. Two distinct processes of electron scattering are known at present: these are scattering by the nucleus and by the extranuclear electrons respectively. These divisions may be further subdivided according as the scattering is elastic or inelastic. No certain instances of inelastic collision of electrons with electrons have yet been reported although Champion[31] has indicated one such possible example. Hughes and West[32] have used the standard electron gun and electrometer arrangement to examine the scattering of electrons of energy 0·002 mv. by the extranuclear electrons of hydrogen. They find complete agreement with the existing theory due to Mott. Very little other work has been done recently on this problem. However, Ruark and Jones[33] have analysed the data obtained earlier by Champion on the scattering of fast β particles by electrons. They show that the suggestion put forward by Jauncey[34] and others that the β particles from radium E might consist of heavy electrons (the existence of which in penetrating radiation seems to be definitely established) is quite untenable. If βc is the velocity of the β particle whose rest mass is R times greater than that of an ordinary electron, then representing $(1-\beta^2)^{-\frac{1}{2}}$ by γ, on Jauncey's theory $R\gamma = \gamma_m$, where γ_m corresponds to the upper limit of the β-ray spectrum and has a value of 3·44 for radium E. Assuming strict conservation of momentum and energy and that the incident β particle has a rest mass Rm_0, the observed value of $R\gamma$ lies between 1·65 and 2·74, most of the values being below 2, whereas Jauncey's theory predicts the constant value 3·44 for this product. If further confirmatory evidence is required we may note that the diffraction pattern obtained by Curran[35] when β particles from radium E are diffracted by Al and Au are just what one would expect from theory, treating the β particles as ordinary electrons.

On the other hand, several papers have appeared on the scattering of fast β particles by nuclei, some of the results being rather divergent in character. First confining the discussion to elastic scattering we may note that Stepanova[36], continuing previous work of Skobeltzyn and Stepanova, reports the scattering of β particles by nitrogen nuclei to be about twice that to be expected on Mott's theory for electrons with energies between 0·2 and 1·1 mv. and as much as 10 to 30 times the theoretical number between 1·5 and 3 mv. The source of electrons used was radium C, and the β rays, selected by a thin canal, were bent through 90° by a magnetic field. In this manner, direct action of the γ rays on the expansion chamber was reduced and velocity selection was obtained. The results in the region of lower energy are in disagreement with earlier work of Champion[37] who examined the collisions of β particles from radium E with nitrogen nuclei. A total track length of 875 m. within the limited energy interval 0·6 to 1 mv. was examined, 201 collisions resulting in deviation through more than 20° being observed; the number predicted by Mott's theory was 236. Similar approximate agreement with theory has been obtained by Zuber[38] who has examined the scattering of β particles by argon nuclei in the energy interval 1·7–2 4 mv. From a total track length of 350 m., 115 collisions

with an angle of scattering greater than 20° were obtained, compared with 138 predicted by theory.

On the other hand, for heavier nuclei, Klarmann and Bothe[39] have reported that the fraction of electrons with energies in the interval 0·5 to 2·5 mv. scattered by krypton and zenon nuclei is only one-fifth of that to be expected theoretically. This anomaly has recently been confirmed by Champion and Barber[40] who find that the fraction scattered by mercury nuclei is only about one-seventh of that indicated by Mott's formula for the energy interval 0·6 to 1·1 mv. The results may perhaps best be summarized as in table 2. Except for workers (1) and (2), the expansion chamber was used, as in the present state of experimental technique this is the only method by which it is possible, in individual cases, to distinguish with certainty between nuclear and electronic scattering.

Table 2

Workers	Scattering nuclei	Energy range mv.	Angular range	Observed / Theoretical
(1) Neher	Al, Au	0·15	90–172°	1·3
(2) Chadwick and Mercier	Al, Cu, Ag, Au	0·3 (mean)	20– 40°	1·4
(3) Stepanova	N	0·2–1·1	20–180°	1·6
(4) Champion	N	0·5–1·1	20–180°	0·9
(5) Champion and Barber	Hg	0·6–1·1	20–180°	0·15
(6) Klarmann and Bothe	Kr, Xe	0·5–2·5	40–180°	0·2
(7) Stepanova	N	1·5–3·0	20–180°	10–30
(8) Zuber	A	1·7–2·4	20–180°	0·8
(9) Oppenheimer and Fowler[41]	Pb	5–17	>10°	1·0

As far as it is possible to draw any general conclusions at present the position would seem to be as follows:

(1) For light elements and all energies the scattering between 20° and 40° is in approximate agreement with theory.

(2) For the lighter elements, except for the results of Stepanova, the scattering at all angles and energies is in approximate agreement with theory.

(3) For heavy elements the scattering is in approximate agreement with theory at energies greater than about 3 mv.

(4) For heavy elements the scattering for energies between 0·6 and 2·5 mv. appears to be much less than that predicted by theory.

The theory is represented by the formula due to Mott which is

$$\sigma_\theta^{180} = \frac{\pi Z^2 r_0^2}{\gamma^2\beta^4}\left[\cot^2\frac{\theta}{2} + 2\beta^2 \log \sin\frac{\theta}{2} + \frac{2\pi\beta Z}{137}\left(\sin\frac{\theta}{2} + \operatorname{cosec}\frac{\theta}{2} - 2\right)\right.$$
$$\left. + \text{higher terms in } \left(\frac{Z}{137}\right)^2, \left(\frac{Z}{137}\right)^3, \text{etc.}\right],$$

where σ_θ^{180} is the effective cross-section for scattering through an angle greater than θ, Z is the atomic number of the scattering nucleus, $r_0 = e^2/m_0c^2$, where e is the electronic charge, m_0 is the rest mass of an electron, and c is the velocity of light. Of the remaining quantities, $\beta = v/c$, where v is the velocity of the incident electron and $\gamma = (1-\beta^2)^{-\frac{1}{2}}$.

This theoretical work is, of course, based on Dirac's relativistic wave-equation for an electron, and it appears to be quite impossible that any slight modification of existing theory can be made to account for a divergence of the magnitude found experimentally. To account for this discrepancy in the region of 1 mv., Champion and Barber have suggested the existence of a repulsive force, superposed on the ordinary Coulomb force during the interaction of electrons of this energy with heavy nuclei. Whether the invention of a new force is justified will require considerable further experiment; if it exists it is clear that it must be confined to encounters of electrons and nuclei since, as we have seen, quantum electrodynamics can be made to account adequately for the scattering of heavy particles by nuclei. The important distinction between light and heavy particles of the same energy is however, that the former must have relativistic treatment at energies in the region of 1 mv., whereas the latter are adequately accounted for by Schrodinger's non-relativistic wave-equation.

We should mention here that Richter[42] has examined the double scattering of electrons of energies up to 0·12 mv. at 90° by gold foils. The results are in complete disagreement with theory, no polarization being observed, whereas 16 per cent of polarization is to be expected from the theory of Mott, based on Dirac's relativistic wave-equation.

Considering next the state of knowledge on inelastic scattering, there is general agreement between all experimenters that for a certain energy range the observed value with heavy elements is considerably greater than that to be expected theoretically, the theory being that due to Bethe and Heitler, based on Dirac's relativistic wave equation. Thus, Klarmann and Bothe found 3 to 5 times the expected amount in krypton and xenon in the energy range 0·5 to 2·5 mv., and Leprince-Ringuet[43] observed 5 to 10 times the theoretical number for β particles of 1 to 3 mv. energy scattered by argon. More recently, Champion and Barber report about 6 times the expected amount of inelastic scattering of 1 mv. electrons by mercury nuclei, while Laslett and Hurst[44] find 30 times the theoretical number for β particles of 1·5 to 4·5 mv. incident on lead nuclei. However, for 2 to 6 mv. electrons scattered by carbon and lead nuclei, Turin and Crane[45] report that the number of inelastic collisions is not very much greater than the theoretical value; further experiment with electrons of energy up to 11 mv. confirmed the indication that the discrepancy was small but real with lead. With carbon the agreement with theory was good. Later experiments of Ruhlig and Crane[46] indicate that there is a great increase in energy loss as the energy of the incident electrons is raised from 0·5 to 1 mv., the number exceeding that to be expected theoretically by a factor of 2 to 4 for lead. On the other hand, the excess of the experimental value above the theoretical value at high energies is again found to be relatively small. This moderate agreement with theory is confirmed by Oppenheimer and Fowler for the energy loss of 9 and 13 mv. electrons and positrons during encounters with lead nuclei. Taken as a whole the evidence is strong that the number of instances of large energy loss found experimentally for electrons between 0·75 and 2·5 mv. is much greater than present theory would allow. For energies lower and greater than this region, however, there is moderately good agreement with the theoretical value for inelastic collisions,

although there still seems to be a slight excess amounting in some instances to 50 per cent in the high energy region.

Considering the lighter elements we may observe that Skobeltzyn and Stepanova[47] found about 100 times the number of inelastic collisions to be expected in nitrogen for β particles of energy between 1·5 and 3·0 mv. However, for the same element, Champion reports that he could find no examples of inelastic collision in 875 m. of track using β particles in the energy range 0·6 to 1·1 mv.; his results receive indirect confirmation from the experiments of Curtis[48] on the energy loss of positrons and electrons of energies 0·3 to 1·5 mv. traversing thin aluminium foils. The energy loss by the two types of electrons was about the same and in agreement with theory in both cases. Further work would appear necessary before any conclusions can be drawn concerning the behaviour of light nuclei. Even with heavy nuclei agreement is not complete since while Klarmann and Bothe found no correlation between angle of deflexion and magnitude of energy loss, Champion and Barber report that for mercury the two quantities generally increase together.

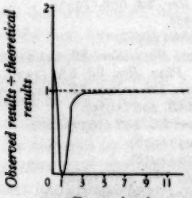


Figure 2(*a*). Elastic.

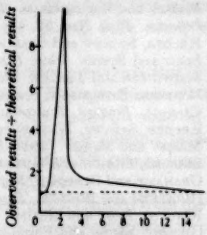


Figure 2(*b*). Inelastic.

It would be a mistake to conclude that the inelastic scattering of the electrons is necessarily accompanied by the emission of the lost energy in the form of electromagnetic radiation. In this connexion the experiments of Arcimovič and Chramov[49] are of interest. A longitudinal magnetic field is used to focus the β rays of energy 2 mv. from radon and examination is made of the production of γ rays when these electrons impinge on solid targets. Using C, Al, Cu, Sn and Pb as the bombarded elements and a counter at 50° to the incident electron beam, the intensity of the γ rays emitted is found to be proportional to the atomic number. This behaviour is in qualitative agreement with the theory of Bethe and Heitler. The variation of the intensity of the γ rays for incident electrons of energies ranging from 0·67 to 3 e.v. was also found to be in good agreement with the theory, the latter predicting 2·9 per cent for the radiation loss in Cu at 2·2 e.v., the observed value being $2{\cdot}8 \pm 0{\cdot}6$ per cent. The possibility that the excess energy may be carried off by some non-ionizing particle, such as a neutrino, during a good proportion of the inelastic collisions of electrons with nuclei must therefore be considered seriously.

In figures 2(*a*) and 2(*b*) are shown graphically the results for the elastic and inelastic scattering of electrons by heavy elements. It would seem that the occur-

rence of a marked trough in the former and a marked hump in the latter for energies in the region of 1–3 mv. must have some important significance. However, the general conclusion at present is simply that the elastic and inelastic scattering of electrons of energy between 0·75 and 2·5 mv. by heavy nuclei is quite different from that predicted by existing theory.

REFERENCES

(1) Bohr. *Nature, Lond.*, **137**, 344 (1936).
(2) Hafstad, Heydenburg and Tuve. *Phys. Rev.* **53**, 239 (1938).
(3) Breit, Condon and Present. *Phys. Rev.* **50**, 825 (1936).
(4) Eddington. *Proc. Roy. Soc.* A, **162**, 155 (1937).
(5) Born. *Nature, Lond.*, **141**, 327 (1938).
(6) Tuve, Heydenburg and Hafstad. *Phys. Rev.* **50**, 806 (1936).
(7) Ochiai. *Phys. Rev.* **50**, 1221 (1937).
(8) Schiff. *Phys. Rev.* **52**, 149 (1937).
(9) Primakoff. *Phys. Rev.* **52**, 1000 (1937).
(10) Heydenburg and Roberts. *Phys. Rev.* **53**, 922 (1938).
(11) Dee and Gilbert. *Proc. Roy. Soc.* A, **163**, 265 (1937).
(12) Massey and Buckingham. *Proc. Roy. Soc.* A, **163**, 281 (1937).
(13) Bonner. *Phys. Rev.* **52**, 685 (1937).
(14) Kruger, Shoupp and Stallmann. *Phys. Rev.* **52**, 678 (1937).
(15) Breit and Stehn. *Phys. Rev.* **52**, 396 (1937).
(16) Schwinger and Teller. *Phys. Rev.* **52**, 286 (1937).
(17) Halpern, Estermann, Simpson and Stern. *Phys. Rev.* **52**, 142 (1937).
(18) Carroll, Powers, Berger and Dunning. *Phys. Rev.* **53**, 680 (1938).
(19) Kruger, Shoupp, Watson and Stallmann. *Phys. Rev.* **53**, 1014 (1938).
(20) Massey and Mohr. *Proc. Roy. Soc.* A, **148**, 206 (1935).
(21) Seaborg, Gibson and Grahame. *Phys. Rev.* **52**, 408 (1937).
(22) Grahame and Seaborg. *Phys. Rev.* **53**, 795 (1938).
(23) Hudspeth and Bonner. *Phys. Rev.* **53**, 928 (1938).
(24) Kikuchi, Aoki and Husimi. *Nature, Lond.*, **141**, 328 (1938).
(25) Gilbert, Smith and Fremlin. *Nature, Lond.*, **139**, 796 (1937).
(26) Goldhaber and Briggs. *Proc. Roy. Soc.* A, **162**, 127 (1937).
(27) McCarthy. *Phys. Rev.* **53**, 30 (1938).
(28) Krammer. *S.B. Akad. Wiss. Wien.* **146**, 71 (1937).
(29) Brubaker. *Phys. Rev.* **53**, 921 (1938).
(30) Gerthsen. *Phys. Z.* **38**, 833 (1937).
(31) Champion. *Proc. Roy. Soc.* A, **163**, 630 (1938).
(32) Hughes and West. *Phys. Rev.* **52**, 43 (1937).
(33) Ruark and Jones. *Phys. Rev.* **53**, 454 (1938).
(34) Jauncey. *Phys. Rev.* **53**, 319 (1938).
(35) Curran. *Phil. Mag.* **24**, 953 (1937).
(36) Stepanova. *Phys. Z. Sowjet.* **12**, 550 (1937).
(37) Champion. *Proc. Roy. Soc.* A, **153**, 353 (1936).
(38) Zuber. *Helv. Phys. Acta*, **11**, 370 (1938).
(39) Klarmann and Bothe. *Z. f. Phys.* **101**, 489 (1936).
(40) Champion and Barber. *Proc. Roy. Soc.* A (in the press).
(41) Oppenheimer and Fowler. *Phys. Rev.* **53**, 928 (1938).
(42) Richter. *Ann. Phys., Lpz.*, **28**, 533 (1937).
(43) Leprince-Ringuet. *Ann. de Physique*, **7**, 5 (1937).
(44) Laslett and Hurst. *Phys. Rev.* **52**, 1035 (1937).
(45) Turin and Crane. *Phys. Rev.* **52**, 610 (1937).
(46) Ruhlig and Crane. *Phys. Rev.* **53**, 618 (1938).
(47) Skobeltzyn and Stepanova. *Nature, Lond.*, **137**, 456 (1937).
(48) Curtis. *Phys. Rev.* **53**, 986 (1938).
(49) Arcimovič and Chramov. *C.R. Acad. Sci. U.R.S.S.* **7**, 415 (1938).

II. COSMIC RAYS

By W. HEITLER, H. H. Wills Physical Laboratory, University of Bristol

§ 1. INTRODUCTION

The last few years have been a time of rapid and decisive progress in our knowledge of cosmic rays. This has been due to the development of the experimental technique as well as to the progress made in the theoretical understanding of the cosmic-ray phenomena.

To mention only some of the important experiments which have been made recently: It has been possible to measure the intensity of cosmic rays at very great heights during balloon flights extending practically up to the top of the atmosphere. Together with the measurements in deep lakes and mines, we have now a full account of the cosmic-ray intensity ranging from nearly the outside of the earth down to 800 m. below sea level.

One of the vital questions in the problem of cosmic rays was that of the energy loss of cosmic-ray particles in traversing matter. This question could be brought to an experimental decision by a refinement of the technique used in measuring the curvature of tracks in a cloud chamber. It is now possible to measure the energies of cosmic-ray particles, and, therefore, also the energy loss, with great accuracy for energies up to several thousand million ev.

On the other hand, the situation could be very much clarified from the theoretical side. In fact an application of the ordinary (relativistic) quantum theory to very fast electrons has revealed the fact that in the region of cosmic-ray energies the behaviour of electrons and light quanta is entirely different from the behaviour at energies obtained from radioactive or artificial sources. Indeed, the theory leads exactly to the type of phenomena observed in cosmic rays. One of the most striking facts—the formation of showers—has found a simple and natural explanation by the laws of ordinary quantum theory. Further investigations have then shown that all the phenomena connected with the so-called soft component of cosmic radiation can be understood more or less quantitatively on the basis of this theory and that the laws of quantum theory are valid for electrons up to the highest energies known at present.

There are, however, facts which could not be explained in this way. The extremely great penetrating power of some of the cosmic-ray particles (the hard component) is in direct contradiction to the quantum theory if the latter is assumed to be valid for electrons. Since this was verified both by direct experiments and by the success with which the theory was applied to the soft component, the only possible conclusion was, that the penetrating particles were not electrons. More recent experiments have in fact shown that we have to deal here with a new type of particle having a mass between that of an electron and proton, e.g. of the order of 100 to 200 electron masses, called the "heavy electron".

The heavy electron discovered in cosmic rays has already proved to be a particle of fundamental importance. It is intimately connected with the short range forces acting between a neutron and proton and also with the anomalous magnetic moments of these particles. It bears the same relation to the nuclear forces as light quanta bear to the electric forces between two charged particles. We shall discuss the connexion between the nuclear properties and the heavy electron in the last chapter of this report. Although it was possible to build up a qualitative theory of the properties of the nuclear particles based on the heavy electron, this theory is still unsatisfactory in many respects and our knowledge about the heavy electron is still very limited. On the other hand, there can be no doubt about the fact that this particle holds one of the key positions for the future nuclear physics and quantum electrodynamics, and that further investigations of its properties will contribute substantially to the final solution of these problems. We believe it will mainly be for experiment to take the lead at present.

§ 2. FUNDAMENTAL EXPERIMENTAL RESULTS

Soft and hard component. In the earlier experiments the intensity of cosmic radiation was measured by an ionization chamber. This method is now usually replaced by an arrangement of two (or more) Geiger-Müller counters placed in a vertical plane and simultaneous discharges of the counters are recorded. In this way the number of single charged particles coming from the vertical direction (or any other given direction) is counted, irrespective of the nature and energy of these particles. It will be seen, however, that in cosmic radiation also neutral particles occur (light quanta, perhaps neutrons and other massive neutral particles) which are not recorded by this method.

In order to gain some knowledge about the properties of the cosmic-ray particles we consider their penetrating power in matter. This can be measured in the following way: A block of some absorbing material (lead, aluminium) with varying thickness is placed between the counters, and the number of particles passing through the block is measured as a function of the thickness.

Auger, Ehrenfest and Leprince-Ringuet[5] were the first who measured in this way the penetrating power of cosmic-ray particles in various materials and at different heights above sea level. The result is given in figure 1 for lead at a height of 3500 m. We see that the number of cosmic-ray particles decreases very rapidly within the first 5 cm. of lead. For larger thicknesses it decreases only very slowly. A certain fraction (about 50 per cent) of the particles is therefore absorbed in 5 cm. of lead, whereas the rest have a much larger penetrating power, up to 1 m. of lead and more. At sea level the absorbable part is only 20 per cent and decreases further below sea level[41]. It was from this result that Auger, Ehrenfest and Leprince-Ringuet first concluded that the cosmic-ray particles can be divided into two groups: a soft group (at sea level 20 per cent) with a penetrating power of not more than 5 cm. of lead, and a hard group with a much larger penetrating power. The distinction between these two groups will prove to be fundamental for the discussion

of all cosmic-ray phenomena and indeed we shall see below that the particles constituting these two groups are different in nature: the soft particles are ordinary electrons, whereas the hard particles are heavy electrons with a rest mass of about 100 or 200 electron masses.

Intensity curve in the atmosphere. The same counter arrangement has served for intensity measurements covering the whole range of the atmosphere and below the earth's surface. Regener and Pfotzer[54,55] have succeeded in sending an apparatus recording the countings automatically in a balloon up to nearly the top of the atmosphere. It was found that the intensity increases rapidly with increasing height above sea level and reaches a maximum very near to the top of the atmosphere, at a pressure of 8 cm. Hg. The maximum intensity is 40 times larger than that at sea level. For still larger heights the intensity decreases again and reaches

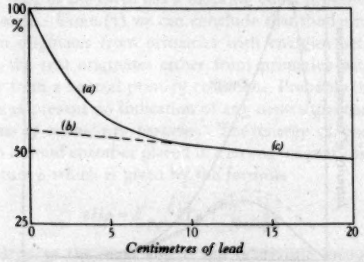


Figure 1. Absorption of vertical cosmic-ray particles (measured on Jungfraujoch). Soft (*a*) and hard (*b*) components.

a certain finite value at the top. The intensity at the top is about one third, certainly less than one half, the maximum intensity.

A certain fraction of the total intensity is due to the hard and soft groups respectively. Measurements (with a lead block of 5–10 cm. between the counters) have so far been carried out only to heights of about 4000 m. The intensity of the soft group increases much more rapidly than that of the hard group. At Jungfraujoch the intensities are about equal (at sea level hard : soft = 4 : 1). In the high layers of the atmosphere the intensity of the hard group is certainly very small compared with that of the soft one.* By extrapolation we find that the ratio hard : soft is only about 1 : 10 at the maximum.

In figure 2 we have plotted the intensity of the soft group only, obtained by subtracting the intensity of the hard group from the total intensity. The fact that this intensity curve shows a maximum near the top of the atmosphere leads to the conclusion that the primary radiation—whatever its nature may be—produces a large number of secondary particles in entering the atmosphere. The secondaries

* By ionization measurements this was also shown by Compton and Stephenson[26].

belong mainly to the soft group and, indeed, the bulk of soft particles are those secondaries produced in the atmosphere.

The intensity of the vertical radiation has also been measured below the surface of the earth. This has been done with a counter arrangement by Ehmert[27] in Lake Constance, and by Barnóthy and Forró[7] in deep mines. Thus we have now a full account of the intensity curve ranging from the top of the atmosphere down to a depth of nearly 1000 m. water equivalent below the top. The intensity decreases steadily (as Ehmert has shown very nearly with the power $h^{1.87}$ of the depth h). At a depth of 100 m. water the intensity is 2 per cent of the intensity at sea level, at a depth of 800 m. water only $\frac{1}{2}$ per cent.

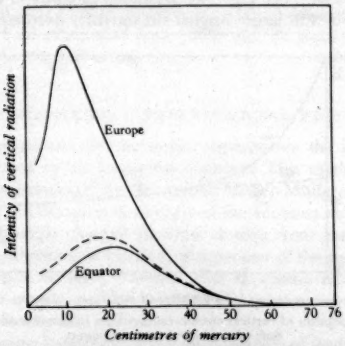


Figure 2. Vertical intensity in the atmosphere in Europe (experiments by Pfotzer) and at the equator. The full line is theoretical, the dotted experimental. (Bowen, Millikan, Neher.)

Latitude effect. It was first found by Clay[25] that the intensity of the cosmic radiation varies with the geographical latitude. The intensity is smaller in the equatorial belt than in Europe. At sea level this variation amounts to only about 20 per cent of the total intensity between European latitudes and the equator. Extensive measurements of the latitude effect covering the whole range of the atmosphere have been carried out by Bowen, Millikan and Neher[19]. The latitude effect is very much larger in the high atmosphere. In figure 2 we have also plotted the intensity at the equator (dotted line). We see that at a pressure of 10 cm. Hg the intensity is only a small fraction of that at 50° latitude. In the high atmosphere the latitude effect is mainly due to the soft component since the hard one is practically negligible there. It is, however, probable that the hard component will also show a certain latitude effect. At sea level both components contribute to the effect.

The effect is caused by the magnetic field of the earth which deflects charged particles coming from the outside of the earth. Although the field is very weak it extends over enormous regions and has, therefore, a considerable influence on

particles with energies even as high as 10^{10} ev.* From the existence of the latitude effect we can conclude that the majority of all incoming rays are charged particles (from which the particles observed at sea level originate as secondaries). The orbits of a charged particle in the magnetic field of the earth have been studied by Lemaître and Vallarta[45]. One of their most important results is the following: charged particles can reach the surface of the atmosphere from the universe only if their energy is larger than a certain critical minimum energy E_λ which depends upon the geographical latitude λ and the direction of their orbit at the earth's surface (or the surface of the atmosphere). For particles reaching the earth in the vertical direction these minimum energies are:

λ	0	50°	90°	
E_λ	2×10^{10}	3×10^{9}	0 ev.	.....(1)

Thus the magnetic field of the earth has a blocking effect preventing slow particles from reaching the earth. From (1) we can conclude that the latitude-sensitive part of cosmic radiation originates from primaries with energies between 3×10^{9} and 2×10^{10} ev. whilst the rest originates either from primaries with energies larger than 2×10^{10} ev. or from a neutral primary radiation. Probably the first alternative is true and there is at present no indication of any neutral primary radiation.

Energy and mass of cosmic-ray particles. The energy of cosmic-ray particles can be measured in a cloud chamber placed in a strong magnetic field H. The tracks show then a curvature ρ which is given by the formula

$$eH\rho = E\sqrt{\left\{1-\left(\frac{mc^2}{E}\right)^2\right\}}, \qquad .....(2)$$

where e is the charge, m the mass and E the relativistic energy of the particle. If e and m are known E can be calculated from the measured curvature ρ. If E is large compared with the rest energy mc^2 the curvature depends only on E and is independent of m. E can then be calculated without knowledge of the mass. If the mass is not known another independent measurement is necessary to determine E and m.† For this purpose a measurement of the specific ionization produced by the particle in the cloud chamber can serve. The specific ionization of a fast particle in standard air depends on the energy in a way shown in figure 3. For small energies it decreases rapidly with increasing energy and reaches a minimum at a kinetic energy equal to the rest energy mc^2 of the particle. For still larger energies the ionization remains more or less constant or, more precisely, it increases logarithmically with increasing ratio E/mc^2. For particles with different rest masses the ionization curves are practically the same, they have merely to be shifted along the energy axis so that the minimum lies at a kinetic energy equal to the rest energy mc^2. The minimum value is the same for all particles with the same charge e. If, therefore, a cosmic-ray track shows an ionization not appreciably higher than

* It is even probable that the magnetic field of the sun has an influence on the cosmic radiation (see Janossy[43] and Epstein[20]).

† The charge of all cosmic-ray particles is $\pm$ the electronic charge. Not a single case has been discovered so far which would favour a different charge. This is concluded from ionization measurements (see below).

the minimum ionization (we call it the normal ionization) we can be sure that the particle has an energy $\gg mc^2$. From a measurement of ρ we obtain then its kinetic energy and at the same time an upper limit for the mass (because the rest energy is then smaller than E). The mass itself cannot then be determined. Only for tracks showing a markedly higher specific ionization than the normal one can m and E be determined.

Most of the cosmic-ray particles have an ionization very near to the normal one. Heavily ionizing tracks are very rare. Thus, for most cosmic-ray particles, we can determine directly their energy but not their mass. The latter can only be obtained from indirect arguments based on the penetrating power, etc. This question, together with the determination of the mass from heavily ionizing tracks, will be dealt with in detail below. Here we anticipate only the result. A certain fraction of the cosmic-ray particles are ordinary electrons. They will be seen to

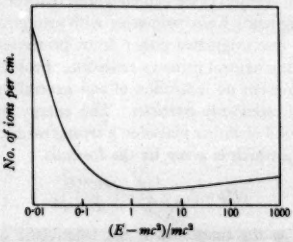


Figure 3. Specific ionization of a fast particle (or specific energy loss by ionization) as a function of the energy. mc^2 is the rest energy of the colliding particle.

constitute the soft component. The hard component consists of particles with a rest mass between 100 and 200 electron masses. Protons (and α-particles) do not seem to occur in cosmic rays in appreciable numbers except as a result of nuclear disintegrations produced by cosmic-ray particles.

The most accurate and most extended energy measurements have been carried out by Blackett and Brode[13,14] (cf. ref. (1)). They were able to measure energies up to 10^{10} ev. According to their results the frequency of occurrence $N(E)\,dE$ of particles with energy between E and $E+dE$ follows approximately an $E^{-2\cdot5}$ law for energies greater than 5×10^9 or 10^{10} ev.:

$$N(E)\,dE = \text{const.}\,\frac{dE}{E^{2\cdot5}}, \qquad \ldots\ldots(3)$$

It will be seen below that this energy spectrum refers mainly to the hard component. For energies less than 5×10^9 ev. it is very difficult to determine the energy spectrum because those particles are frequently produced as secondaries in the walls of the chamber or in the surrounding materials (magnet), a fact by which the primary energy spectrum is, of course, very much falsified.

Further experiments, in particular on the energy loss of cosmic-ray particles in traversing matter, will be discussed below.

Showers. We have seen in the preceding section 2 that the primary cosmic-ray particles are capable of producing a large number of secondaries in passing through the atmosphere. The same is the case if a cosmic-ray particle passes through a sheet of any heavy material (lead). It happens then that a single particle creates a whole bundle of secondaries. This phenomenon is called a *shower* and was first discovered by Blackett and Occhialini[12] on direct Wilson chamber photographs. These showers were produced in the metal walls of the chamber itself. A closer study of the phenomenon shows that showers are also frequently produced by non-ionizing rays (light quanta).

The shower phenomenon can be studied more quantitatively by means of Geiger-Müller counters arranged, for instance, as shown in figure 4*a*. The shower is created in the metal plate and simultaneous discharges (coincidences) of all

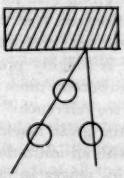


Figure 4*a*.

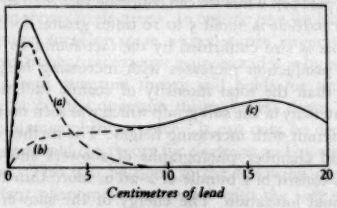


Figure 4*b*.

Figure 4. Rossi curve. 4*a*, experimental arrangement. 4*b*, number of coincidences as a function of the thickness (schematic). (*a*), (*b*), contribution of the soft and hard components respectively.

three counters are recorded. It is necessary that at least two particles (or more if more counters are used) emerge simultaneously from the plate in order to produce such a coincidence. Rossi[56] was the first who measured the number of coincidences as a function of the thickness of the shower producing plate. In figure 4*b* we give a typical transition curve for lead[58,34]. It has a maximum at about 15 cm. of lead and then decreases rapidly. After 5 cm. an almost constant value is reached. At 15 to 20 cm. the curve has a flat second maximum. The shape of the Rossi curve shows that the shower production is a complex phenomenon which is certainly due to various types of processes. This was also to be expected since the shower producing radiation consists of two groups with very different penetrating powers.

The curve in figure 4*b* may be considered as composed of two parts (*a*) and (*b*). (*a*) is responsible for the first maximum and decreases to zero within 3 to 5 cm. of lead. This part of the curve is certainly due to showers produced by the soft component, which was seen in § 1 to have a penetrating power of a few cm. of lead. If such a radiation produces secondaries with a penetrating power of the same order of magnitude, i.e. 1 or 2 cm. of lead, we obtain for the number of showers

just a curve of the type (*a*). At first the number of showers increases with increasing thickness owing to the fact that the probability for producing a shower increases with the amount of material traversed. This, however, is only the case as long as neither the primary nor the secondary radiation is absorbed appreciably. A maximum is reached at a thickness corresponding to the mean penetrating power of these two radiations. For larger thicknesses both the primary and the secondary radiation are absorbed again and the curve decreases rapidly.

The second part (*b*) of the Rossi curves rises rapidly (within 1 or 2 cm. of lead) to an almost constant value. It is certainly due to a hard shower-producing radiation which yields soft secondaries with a penetrating power of 1 or 2 cm. of lead.* We shall discuss this part of the curve, together with the second maximum, more in detail below (§ 7).

We have seen above that 80 per cent of all particles at sea level are hard ones. In figure 4 the constant value of part (*b*) is only about half the maximum value of the soft part (*a*). Thus we can conclude that the rate of shower production per single primary particle is about 5 to 10 times greater for the soft group than for the hard one. This is also confirmed by the fact found by many observers that the rate of shower production increases with increasing height above sea level much more rapidly than the total intensity of cosmic radiation. It is roughly proportional to the intensity of the soft group which was seen on p. 364 to become more and more preponderant with increasing height. For further details see reference[38].

Cloud chamber photographs of showers show that the great majority of all showers consist of a bundle of 2–20 or more thin tracks with an ionization near to the normal ionization. The energy of the shower particles ranges in most cases between 10^7 and 10^8 ev. The mass of these particles is, therefore, small and it is certain that most of the shower particles are ordinary positive and negative electrons. The electrons emerge from a centre in more or less the same direction and the angular spread of a shower is—except for very large showers—quite small.

In some cases very large showers with 100–1000 particles have been observed. They are usually called "bursts". The phenomenon had already been discovered a long time ago by Hoffmann[42] in an ionization chamber. The ionization due to cosmic rays was recorded as a function of time. Occasionally a sudden very large increase of ionization ("Stoss") was observed. If this sudden burst of ionization was due to particles with normal ionization, a very large number of particles (100 to several thousand) must have entered the chamber simultaneously. It is now quite certain that these bursts are identical with the very large showers observed in the cloud chamber, and it has been shown by Bøggild[18] and Carmichael[34] that the Rossi curve for bursts is very similar to that for small showers, the only difference being that the contribution from the hard component is comparatively larger.

In some cases showers of a type quite different from the bundle showers have been observed[22]. In contrast to the latter they contain heavily ionizing tracks (accompanied by thin tracks) and these particles are emitted in all directions from

* The part (*b*) of the shower curve can be measured more easily below sea level where the hard component is very much more preponderant[6].

a point. The ionization of the thick tracks shows that these particles are in all probability protons (at least in most cases). It is probable that we have to deal here with nuclear disintegrations produced by cosmic rays. These showers can also be observed directly in a photographic plate exposed for some time to cosmic rays [17]. The shower produced in the emulsion of the plate appears as a little star consisting of a number of proton tracks emitted all from a single point (thin tracks are not recorded by this method). We discuss this type of shower in the last section.

§ 3. THEORY OF THE PASSAGE OF FAST ELECTRONS THROUGH MATTER

From the experiments discussed in the preceding section it seems very probable that at least a certain fraction of the cosmic-ray particles are ordinary positive and negative electrons (for instance, the shower particles). For a theoretical understanding of the cosmic-ray phenomena it will thus be the natural way to try to apply the ordinary (relativistic) quantum theory to electrons. In cosmic rays we have, however, to deal with particles having energies 10 to 10,000 times larger than the highest energies available from artificial or radioactive sources and it is by no means a priori certain whether our present quantum theory still holds for those high energies. The question whether or not the quantum theory breaks down for high energies is, of course, one of fundamental importance. It can only be decided by working out in detail the consequences of the theory for electrons and by comparing the results with the experiments. This will be done in the following section. It will be seen that all the important phenomena connected with the soft radiation can be understood on grounds of the quantum theory of the electron and that the latter holds for energies up to the highest known at present.

Energy loss of fast electrons. A fast particle passing through matter loses its energy mainly in two ways:

(*a*) In passing through the electronic shell of an atom energy is transferred to one of the electrons by an inelastic collision. This is the familiar process of ionization by fast particles. The mean energy transferred to the electron is practically independent of the energy of the primary particle. The curve shown in figure 3 giving the ionization of a fast particle stands, therefore, approximately also for the mean energy lost by the primary particle. The energy loss per cm. path of a fast electron with an energy much greater than mc^2 is given by the following formula:

$$\left(-\frac{\partial E}{\partial x}\right)_{\text{coll.}} = NZ2\pi\left(\frac{e^2}{mc^2}\right)^2 mc^2 \log\frac{E^3}{2mc^2I^2Z^2}, \qquad \ldots\ldots(4)$$

where N is the number of atoms per cm^3, Z the atomic number, I a certain mean ionization energy, viz. $I = 13{\cdot}5$ ev. The specific energy loss is proportional to the number of electrons NZ per unit volume, e.g. to Z.

(*b*) If an electron passes through the electric field of an atom it is deflected, which is equivalent to a certain acceleration of the electron. Thus radiation will always be emitted in such a collision. The process is known as ordinary emission of a continuous x-ray spectrum. For cosmic-ray energies it plays the most important

part for the energy loss of the electron. The theory of this process has been developed by Bethe and Heitler[5, 37]. We confine ourselves to a brief summary of the results.*

An electron moving with cosmic-ray energies suffers in general only very small deflections in passing through an atom, but even a small deflection represents, for a very fast particle, a large acceleration and the emission of light becomes, therefore, very probable. The light quantum is emitted in practically the same direction as the electron, e.g. in the forward direction. The theory shows now that there is a large probability for the emission of a light quantum having an energy $h\nu$ comparable with the energy E of the primary electron itself. If we express the energy of the light quantum as a fraction of E, $\alpha = h\nu/E$, the probability of an electron losing the fraction α of its energy is, for high energies, independent of E itself. The mean energy lost by the electron in passing through an atom increases, therefore, proportionately to the energy E of the electron. The frequency distribution of the energy lost is nearly uniform, thus, for instance, three-quarters of the energy is lost in the form of quanta having each an energy larger than $E/4$.

The total mean energy loss of an electron by radiation per cm. path is given by the formula (for very high energies)

$$\left(-\frac{\partial E}{\partial x}\right)_{\text{rad.}} = NZ^2\left(\frac{e^2}{mc^2}\right)^2 \frac{E}{137}\left[21 - \tfrac{4}{3}\log Z\right]. \qquad \text{.....(5)}$$

It has, however, to be born in mind that this energy is not lost steadily, but in the form of a few hard quanta with energy $h\nu$ each comparable with E itself. The energy lost by a particular electron in traversing a thin plate of matter fluctuates therefore considerably about the mean loss given by equation (5). Furthermore, in contrast to the energy loss by ionization, (5) is proportional to the square of the nuclear charge Z.

Since $(\partial E/\partial x)_{\text{rad.}}$ increases $\sim E$ there will be a certain critical energy E_c for which $(\partial E/\partial x)_{\text{coll.}} = (\partial E/\partial x)_{\text{rad.}}$. For energies $E > E_c$ the energy loss by radiation will be the only important one, all other modes of energy loss being then negligible. This critical energy depends on Z and is higher for light elements than for heavy ones.

Comparing (4) with (5) we find easily that E_c lies at

Air and water	Al	Pb	
$E_c = 1{\cdot}3 \times 10^8$	6×10^7	1×10^7 eV.	.....(6)

Thus, cosmic-ray electrons will practically lose their energy only by radiation.

Absorption of light quanta. The absorption laws of light quanta are very similar to those of fast electrons. At low energies light quanta are absorbed by the photoelectric effect and by Compton scattering. At cosmic-ray energies, the only important mode of absorption is that by the creation of a pair of a positive and negative electron. There is again a critical energy E_c so that for $E > E_c$ all other processes can be neglected whilst for $E < E_c$ the Compton scattering (and for still smaller energies the photoelectric effect) is more important than pair production. E_c is

* Cf. also the article by Co'lie, *Progress Reports* (1937).

about the same as for electrons, equation (6). The absorption coefficient for hard light quanta due to pair creation is given by the formula (for very high energies)

$$\tau = N \frac{Z^2}{137} \left(\frac{e^2}{mc^2}\right)^2 [16{\cdot}1 - \tfrac{4}{3} \log Z]. \qquad \ldots\ldots(7)$$

Thus, the absorption coefficient does not depend on the frequency at all and is nearly proportional to Z^2. (The absorption coefficient due to Compton scattering decreases $\sim 1/\nu$ and is proportional to Z.)

The formulae (5) and (7) show that fast electrons and hard light quanta are highly absorbable. Their penetrating power does not, or practically not, increase with increasing energy in contrast to the behaviour at lower energies. In normal air a cosmic-ray electron of 100×10^6 ev. has a mean range of only 200 m. A light quantum of the same energy has an absorption coefficient of $0{\cdot}2 \times 10^{-4}$ cm^{-1} (mean range of 500 m.) In lead these figures are 1·5 cm. and 1 cm^{-1} respectively.

At first sight it seems that this theory leads to a penetrating power for electrons far too small to account for the observed behaviour of cosmic particles. But this difficulty is only an apparent one and we shall see in § 4 that the theory is fully confirmed by the cosmic-ray facts.

§ 4. THEORY OF CASCADE SHOWERS

We have seen that a fast electron loses its energy in matter within a very small range. In order to apply the theory to the actual passage of cosmic-ray electrons and light quanta through thick layers we have to consider the way in which the energy is dissipated more in detail. According to § 3 the energy lost by the electron is not divided into a very large number of small quanta (as is the case for the energy lost by ionization) but reappears in the form of a few hard light quanta with energies each comparable with the energy of the primary electron. All these quanta are practically emitted in the forward direction. They carry a large fraction of the energy of the original electron somewhat farther through the layer until they are absorbed (a small fraction of the energy is lost in the form of soft quanta emitted in all directions). The hard light quanta are absorbed by creation of an electron pair, the whole energy of each quantum being divided up between the two pair electrons. Each of these pair electrons will, therefore, have an energy, though of course smaller than, but still comparable with, the primary energy E_0. At this stage of the process we have instead of one primary electron with energy E_0 a number of positive and negative electrons with energies still comparable with E_0, although, of course, a certain fraction of E_0 is really lost in the form of soft quanta which cannot produce pairs.

Each of the secondary electrons behaves now in a similar way to the primary electron, e.g. it will emit hard light quanta which produce pairs and so on. This results in a further multiplication of the number of electrons together with a decrease of the energy for each particle. The multiplication takes place as long as the energy of the electrons and light quanta is larger than the critical energy E_c. As soon as this energy is reached no further multiplication can take place; the electrons then

lose their energy steadily by onization and the light quanta by Compton scattering. The larger the primary energy is, the more stages are possible and the higher will be the number of electrons produced. Each stage, however, requires a certain thickness of the absorbing layer, which is about equal to the average range which a fast electron travels before its energy decreases by, say, half the primary energy. For thin layers only few stages are possible. In very thick layers the process will result in a complete degradation of energy, all electrons and light quanta being ultimately absorbed. There will be an optimum thickness l_m at which the number of electrons is a maximum. If we interrupt the process at a thickness near to l_m and observe the outcoming electrons, say, in a Wilson chamber, we shall find exactly what we have called a *shower*. It is this multiplicative process which is responsible at least for the majority of the showers observed in cosmic radiation.

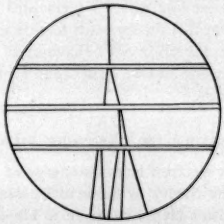


Figure 5. The cascade process for shower production (schematic). The incident particle is multiplied by three successive lead plates.

Recent experiments by Trumpy[61] and Fussel[33] have shown this directly. A number of very thin lead plates were placed in a cloud chamber some distance apart from each other. It can then be seen how the multiplication process is progressing from plate to plate. We give a schematical picture of such a cascade process in figure 5. The electron on the left-hand side produces a hard light quantum in the second plate which produces then a pair in the third plate.

The cascade shower process has been treated quantitatively independently by Bhabha and Heitler[9] and by Carlson and Oppenheimer[23]. In the following table 1 we give some of the numerical results. The mean number of electrons (of both signs) $\bar{Z}$ contained in a shower which is produced by a primary electron with energy E depends on E, the material in which the shower is produced and the thickness of the layer. It is convenient to introduce for the thickness a certain unit length depending on the material in question, and representing roughly the distance which an electron travels before it loses its energy by radiation.

The unit thickness is for:

Normal air	Water	Aluminium	Lead	
280 m.	35	6·7	0·36 cm.	.....(8)

Furthermore, we introduce instead of the primary energy E_0 the variable

$$y = \log \frac{E}{E_c}, \qquad \text{.....(9)}$$

where E_c is the critical energy defined in equation (6) which also depends on the material.

If these variables are chosen $\bar{Z}$ is a function of l and y only. In table 1 we give the number of shower electrons as a function of l for various values of y. $\bar{Z}$ increases very rapidly with y. According to this table an electron of 5×10^8 ev. would produce in a lead plate of 1·1 cm. ($l=3, y=4$) on the average 6·4 shower electrons. A primary electron with energy 4×10^9 ev. would produce a maximum number of more than 50 electrons. Thus we see that even very large showers and Hoffmann bursts can be explained by our theory.

In general, most of the shower electrons have energies near to E_c (except for very thin plates), but higher and smaller energies also occur.

Table 1. Total mean number of shower electrons*

y \ l	1	2	3	5	10	18
2	1·2	1·6	1·8	1·4	0·1	0
4	1·9	4·0	6·4	9·6	4·1	0
6	2·7	7·7	15·5	40	54	10
10	4·7	20	57	260	1800	1500

From this table we also see how the mean number of shower electrons depends upon the thickness l. $\bar{Z}$ starts at a value 1 (primary electron) at a thickness $l=0$ and rises to a maximum which lies at a value of l which increases with the primary energy (or y). At large thicknesses all electrons are absorbed and $\bar{Z}$ falls towards zero.

Table 1 gives only the mean number of shower electrons. The actual number of electrons produced fluctuates again largely about the mean value.

From the above considerations it is clear that light quanta are also capable of producing cascade showers in almost exactly the same way as electrons. The first stage of the process is then a pair creation followed by radiation.

§ 5. APPLICATIONS OF THE THEORY

The theory of cascade showers described above has found numerous applications to all sorts of phenomena in cosmic rays. In this report we must confine ourselves to a brief discussion of some of the most important applications.

The Rossi curve. If we plot the number of shower electrons $\bar{Z}$ given in table 1 as a function of the thickness, we obtain a curve which resembles very much the Rossi curve for triple coincidences shown in figure 4*b* part (*a*), if the effect due to the hard component, part (*b*), is subtracted. The theoretical curve for $\bar{Z}$ depends

* The numerical calculations have been extended by Arley[3].

upon the primary energy of the electron falling on the top of the shower-producing plate. The position of the maximum shifts to higher thicknesses for higher primary energies. For a rough orientation we might compare the experimental curve with the curve for $\bar{Z}$ which has the maximum at the same position as the experimental one. This is the case for $y = \log(E/E_c) = 3$ corresponding for lead to a primary energy (for lead $E_c = 10^7$ ev.) of 2×10^8 ev. We may expect, therefore, that most of the shower-producing electrons have about this energy. This will be seen in § 6 to be the case.

Strictly speaking, the experimental curve cannot be compared directly with the curves representing the mean number of shower electrons. In the first place the experimental curve is the result of primary electrons with all different energies.

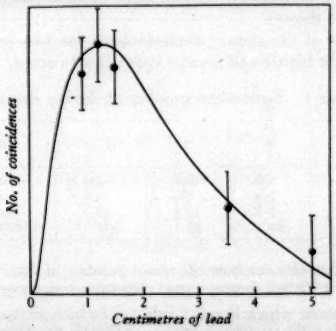


Figure 6. Number of cascade showers containing two, or more than two particles. Theory (full curve) and experimental points by Schwegler. (Experimental arrangement similar to figure 4(*a*), the hard component is eliminated experimentally.)

To take this into account we should have to know the energy spectrum of the shower-producing electrons and light quanta. This is only roughly known so far; we shall discuss it below. In the second place, what is measured by the experimental arrangement shown in figure 4*a* is not the mean number of shower electrons, but the number of showers containing a certain minimum number of particles (in figure 4*a* this number is two). The probability for the production of a shower with, say, more than 2 or 3 particles by a given primary electron can be calculated from the mean number of shower particles, taking into account the fluctuations about this mean number.

Arley[3] has calculated in this way the exact shower curve with the energy spectrum discussed in § 6. The result is shown in figure 6 where the (theoretical) probability for the production of a shower of at least two particles is plotted normalized for one incident electron. The experimental points are obtained by Schwegler[30],* the scale for the latter is chosen so that the height of the maximum

* In this experiment the contribution of the hard component is eliminated experimentally.

is the same as for the theoretical curve (the position of the maximum cannot, of course, be chosen arbitrarily). The position of the maximum and the shape of the curve agree very well with the theoretical one.

In a similar way all the other counter experiments can be treated theoretically. For instance, it is possible to calculate the soft part of the absorption curve[39, 3], figure 1. On the whole the agreement between theory and experiment is very good, but it is always the soft part of the phenomena which is accounted for by the cascade theory, whilst the hard part remains unexplained. The conclusion is, therefore, that the soft radiation consists of electrons and light quanta which behave in full accordance with the quantum theory.

For Hoffmann bursts Euler[30] and Montgomery and Montgomery[40] have discussed the transition curve and have found that the cascade theory can well account for a certain fraction of the bursts (50 per cent in lead of 5 cm. thickness). For those very large showers primary energies of 10^{10} to 10^{11} ev. are necessary and these considerations show, therefore, that the quantum theory is still valid at these energies. On the other hand, the contribution of the penetrating component is more preponderant for bursts than for small showers.

The intensity curve in the atmosphere. The intensity curve shown in figure 2 has a shape very similar to an ordinary Rossi curve in lead. It is therefore likely that it can be explained in a similar way. The maximum is due to the fact that the rays coming from outside the earth produce numerous secondaries in the upper layers of the atmosphere which are then absorbed again at increasing thickness. In order to compare this curve with our theory we must remember that the energies of the incoming particles are restricted by the magnetic field of the earth. Assuming that the incoming rays are ordinary electrons, we take from equation (1) that all these electrons have energies $\geqslant 3\times 10^9$ ev. at a latitude of 50°. Presumably, the bulk of these particles will have energies just above this minimum energy. Since for air the critical energy E_e is 15×10^7 ev. the y value for the primary electrons is $\log(3\times 10^9/15\times 10^7)=3$. According to the results given in table 1, each primary electron produces a total maximum of electrons (including the primary) of 2·5 electrons. The maximum lies at a thickness of $l=3$, i.e. in air 8 cm. Hg, which is exactly what is observed. Also the height of the maximum (about three particles per one primary particle) is in good agreement with the theory although an exact comparison is hardly possible because the number of incident particles is known only approximately from experiments.

The good agreement leaves no doubt about the fact that most of the incident rays are ordinary electrons and that the cascade theory can account for the soft part of the intensity curve, figure 2.

If we consider the intensity curve down to lower heights we find still a considerable number of electrons (quite apart from the hard radiation). If we were to assume only primaries with an energy of 3×10^9 ev. the theory would not yield enough electrons at greater depths below the top of the atmosphere. These electrons must be due to primaries with energies above 3×10^9 ev., which, though few in number, produce many secondaries at great thicknesses. From the observed intensity in

the atmosphere it is then possible to calculate the energy distribution of the primaries. Let $F(E)\,dE$ be the number of primary electrons in the energy interval dE and $Z(l, E)$ the total number of particles produced by one primary particle with energy E at a thickness l, as given in table 1, the observed intensity $T(l)$ is

$$T(l)=\int_{3\times 10^9}^{\infty} F(E)\,Z(l, E)\,dE. \qquad \text{.....(10)}$$

From this equation $F(E)$ has been calculated by Heitler[38] and Nordheim[52] and it has been found that the primary energy distribution is approximately given by the law

$$F(E)=\frac{\text{const}}{E^{2\cdot 5}} \quad (\text{for } E>3\times 10^9 \text{ ev.}). \qquad \text{.....(11)}$$

An experimental test for this law is provided by the latitude effect. In § 2 we have already seen that at the equator all particles which have energies $\leqslant 2\times 10^{10}$ ev. are prevented from reaching the earth. Knowing the primary energy spectrum we are able to predict the intensity curve at the equator. We simply have to cut off the effects of all particles with energy $<2\times 10^{10}$ ev., thus

$$T_{\text{eqn.}}(l)=\int_{2\times 10^{10}\text{ ev.}}^{\infty} F(E)\,Z(l, E)\,dE. \qquad \text{.....(12)}$$

The calculated curve $T_{\text{eqn.}}(l)$ is also plotted in figure 2 (full drawn). The experiments (dotted curve) have been carried out by Bowen, Millikan and Neher[19], and we see that their results are in perfect agreement with our predictions.

Finally, we mention another consequence of the primary energy spectrum equation (11). By the cascade theory and the primary energy spectrum it is not only possible to compute the total number of electrons at each point of the atmosphere, but also their energy spectrum. At sufficiently large thicknesses the result is extremely simple. A primary energy spectrum of the type equation (11) produces at large thicknesses secondaries with energies following the same energy spectrum but for all energies larger than the critical energy E_c. Thus in air the energy spectrum of the electrons is given by

$$f(E)\,dE=\frac{\text{const}}{E^{2\cdot 5}}\,dE, \quad E>1\cdot 5\times 10^8 \text{ ev.} \qquad \text{.....(13)}$$

For $E<1\cdot 5\times 10^8$ ev. $f(E)$ is approximately constant. This law is valid for all heights less than about 5000 m. above sea level.

The energy spectrum (13) cannot be compared directly with the measurements (§ 2) because the measured energy spectrum refers to the total radiation (hard + soft), whereas our calculated energy spectrum is that of the soft component only. It is, however, satisfactory that the two components have, at sea level, similar energy distributions. An indirect experimental evidence for the energy spectrum of the soft component is provided by the distribution in size of large showers and bursts[16,49]. This has been discussed by Blackett[16] and it was found that the energy distribution, equation (13), was in agreement with the experiments. We discuss the energy spectrum again in § 6.

There are a number of other details connected with the soft component which can be explained satisfactorily on the basis of the cascade theory and its applications to the phenomena in the atmosphere, but for these questions we must refer the reader to the original papers.

§ 6. DIRECT ENERGY LOSS MEASUREMENTS AND THE NATURE OF THE HARD COMPONENT

The energy loss curve. As we have seen in the preceding sections a number of cosmic-ray phenomena have found a qualitative and quantitative explanation by the cascade shower process which is a direct consequence of the quantum theory applied to electrons and light quanta of very high energies. This theory, however, leaves no room for an explanation of the hard component. According to the theory an electron can make its effects felt (by producing secondaries) through layers of at most 5 or 10 cm. of lead, 10 m. of water or 10 km. of normal air. The penetrating cosmic rays are found to pass through layers of 1 m. of lead and more and there are still particles found in mines at a depth of 800 m. below sea level. It is clear that these penetrating particles cannot be electrons if for the latter the cascade theory is adopted for all energies. It was in fact from the success of this cascade theory that several authors[38, 52, 50] first came to the conclusion that the penetrating particles are not electrons.

On the other hand, it can also be shown that the penetrating particles cannot be of protonic mass. The ionization of most of these particles is the normal ionization (minimum in figure 3) which shows that their kinetic energy is higher than the rest energy. Deflection experiments in a magnetic field show that these particles have energies down to 2×10^8 ev. Since the rest energy of a proton is $Mc^2 = 10^9$ ev. it is impossible that the penetrating particles have protonic mass. Their rest energy cannot be larger than about 2×10^8 ev. corresponding to 400 electron masses. The only possible assumption seems, therefore, to be that the penetrating particles are neither electrons nor protons but represent a hitherto unknown type of particle.

The argument leading to this decisive conclusion rests on the validity of the radiation formulae (5, 7), for the energy loss of electrons up to extremely high energies. The evidence for the validity of the radiation formulae was collected from the phenomena connected with the soft radiation. Although this evidence is very strong it is only indirect, and it is therefore desirable to find a direct experimental proof for the large energy loss of electrons predicted by the theory on the one hand, and for the different behaviour of the penetrating particles on the other hand.

The specific energy loss of a cosmic-ray particle can be measured in a Wilson chamber by placing a metal plate (for instance, 1 cm. lead) across the chamber. In a magnetic field the tracks traversing the plate show a larger curvature below the plate because the particles have lost energy in the plate. From the difference of the curvatures below and above the plate the energy lost by the particle is measured.

The first measurements in this direction were made by Anderson and Neddermeyer[2], while the most accurate and most extended ones are those by Blackett

and Wilson[15 16,64]; ref. (7). We give their results schematically in figure 7, where the mean specific energy loss per cm. of lead is plotted as a function of the primary energy. The theory would give for electrons a straight line $-(1/E)\,\partial E/\partial x=2$ (dotted line). The full drawn curve represents the mean value of the experimental energy loss for each energy. As mentioned in § 4 the actual energy lost in each particular case fluctuates largely about this mean value. Up to an energy of 2×10^8 ev. the mean energy loss is in exact agreement with the theory. For higher energies the curve drops down rapidly and reaches a value smaller by an order of magnitude than the theoretical loss of an electron. This is not surprising because we already know that there are particles having a much greater penetrating power than electrons would have according to the theory. We now see that these penetrating particles are mainly those which have energies $>2\times10^8$ ev. whereas the soft particles are mainly those with lower energies. It would, however, be surprising

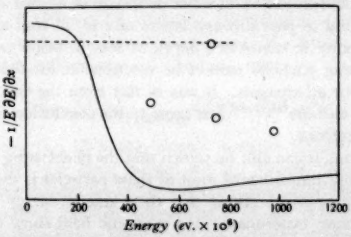


Figure 7. Mean specific energy loss of cosmic-ray particles. Experiments by Blackett and Wilson. The dotted line corresponds to the theoretical value for electrons. In the energy region 2×10^8 ev. the graph shows the existence of two groups of particles: a hard group with a small energy loss represented by the full line and a very small number of soft particles (circles) having the energy loss expected for electrons. For low energies (2×10^8 ev.) all particles are electrons obeying the theory.

if these two groups were distinguished so sharply by their energy and the question now arises whether there are also soft particles with energies $>2\times10^8$ ev. or hard ones with energy less than 2×10^8 ev. This is in fact the case.

In the region $E>2\times10^8$ there have been found a small number of particles having a much larger energy loss than most of the other particles of the same energy. We have marked these particles by a small circle in figure 7. Their energy loss is in agreement with that expected for electrons. Moreover, these particles are almost always associated with showers as is to be expected for electrons of such high energy according to the cascade theory. The number of these highly absorbable particles is comparatively small, not more than 1 or 2 per cent of all particles.

These experiments show directly that in the energy region $E>2\times10^8$ ev. there are two groups of particles which behave differently and which must therefore be of a different nature. There is a small number of highly absorbable particles which behave like electrons according to the quantum theory and which are certainly

electrons. On the other hand, the bulk of particles with $E > 2 \times 10^8$ ev. have a small energy loss. These particles cannot therefore be electrons. We discuss their nature below.

The existence of these few tracks with large energy loss provides also a conclusive proof of the fact that for electrons the quantum theory is valid even for very high energies, and there is no reason to believe that there is any limit to the validity of the theory at all. This we had already concluded from the success of the cascade theory applied to large showers.

In the region $E < 2 \times 10^8$ ev. practically all particles are electrons. In Blackett and Wilson's experiments no particles have been found with considerably greater penetrating power than electrons and energy $< 2 \times 10^8$ ev.* Since, however, the individual energy loss fluctuates largely about the mean value it might be that some of the particles with comparatively small energy loss are in fact penetrating particles.

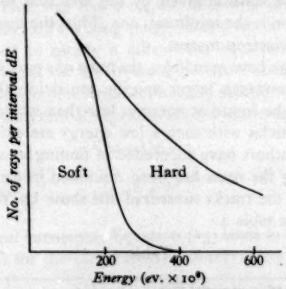


Figure 8. Energy spectra of the soft and hard components (schematic) according to Blackett.

In any case their number is small compared with the number of electrons and, furthermore, they must have lost much more energy than their fellow particles lose in the region of higher energy $> 2 \times 10^8$ ev. Thus it seems that the energy loss of the penetrating particles increases when their energy passes below 2×10^8 ev.

We can now give a schematic picture of the energy distribution of the soft and hard particles separately, figure 8. Most of the hard particles have energies greater than 2×10^8 ev. whereas most of the electrons have energies less than 2×10^8 ev.

The exact forms of the energy spectra of the two components are only partly known at present, and only for the high energy region. The theory shows that the soft component has an energy spectrum given by equation (13) for $E > 2 \times 10^8$ ev. The energy measurements (§ 2) show that a similar energy spectrum holds for the total radiation and therefore also for the hard component for $E > 5 \times 10^8$ ev. The energy spectra for the hard component for $E < 5 \times 10^8$ ev. and for the soft one for $E < 2 \times 10^8$ ev. are not known at present.

* Other authors have found a few heavy electrons with energy less than 2×10^8; see below.

The nature of the penetrating particles. Having seen from direct as well as from indirect evidence that the penetrating particles are neither electrons nor protons, the question arises what the nature of these particles may be. The first question is why these particles lose so little energy. The energy loss of a particle in a given material depends, according to § 3, only on two parameters, the charge e and the mass m. The energy loss by ionization depends only on e; in formula (4) m outside the logarithm is the mass of the extranuclear electrons in an atom and is not the mass of the colliding particle in question. The ionization of the penetrating particles is found to be the same as for electrons and e must therefore be the elementary charge. The energy loss by radiation equation (5) depends also upon the mass m of the colliding particles and is proportional to $1/m^2$. The only way to cut down the large energy loss by radiation is therefore to assume a mass considerably larger than the electronic mass; this suggestion was first made by Neddermeyer and Anderson[50]. An upper limit for the mass is given by the fact that for energies greater than 2×10^8 ev. the ionization is the minimum one. Thus the mass cannot be larger than 2×10^8 ev./c^2, i.e. 400 electron masses.

According to what we have seen in § 2, the mass of a particle can only be measured if tracks are found showing a larger specific ionization than the minimum one. Those tracks can only be found at energies less than 2×10^8 ev. As we have seen above, penetrating particles with such a low energy are extremely rare. In spite of this difficulty several authors have succeeded in finding some of those tracks.* The method of determining the mass has been discussed in § 2. We shall not give the details here. Most of the tracks measured are those by Williams and Pickup[63]. We give their results in table 2.

Table 2

Momentum (unit mc, m = electron mass)	Mass (unit electron mass)
69	220 ± 50
110	430
95	190 ± 60
73	160 ± 30

Thus we see that all particles, with one exception, have the same mass within the experimental errors and this mass is about 100 or 200 electron masses. The new particle has been called by several authors the heavy electron.

The heavy electron is thus the particle constituting the penetrating cosmic radiation. We discuss some of its properties and applications in the following section.

* The first track of this kind with the right interpretation was found by Street and Stevenson[60]. For other tracks cf.[57,51,28].

§ 7. THE HEAVY ELECTRON AND THE NUCLEAR PROPERTIES

Yukawa's theory. The discovery of the heavy electron has already proved to be one of fundamental importance. It seems now quite certain that the heavy electron holds the key position for an understanding of the elementary properties of the nuclear particles proton and neutron. It is well known that between a proton and neutron (and also between two protons and probably between two neutrons) strong forces act with a range of only 3×10^{-13} cm. These forces cannot be of any electromagnetic nature. Furthermore, both particles have an anomalous magnetic moment which, expressed in units of the Bohr nuclear magneton, have the values $\mu_P = 2{\cdot}46$, $\mu_N = -1{\cdot}80$. These magnetic moments have found no explanation in any of the existent theories.

In a very ingenious paper published in 1935, Yukawa[65, 66] proposed a theory of the nuclear forces which was based on the following ideas: Realizing that the electromagnetic forces are always long range forces, he introduced a new type of field which was assumed to satisfy a differential equation leading automatically to short range forces. Whilst the electrostatic potential satisfies Poisson's equation $\nabla^2\phi = 0$ which, for an electron for instance, has the solution e/r, Yukawa's potential is assumed to obey the differential equation (in the static case)

$$\nabla^2\phi + \lambda^2\phi = 0 \qquad \text{.....(14)}$$

with the solution

$$\phi = \frac{g}{r} e^{-\lambda r}. \qquad \text{.....(15)}$$

λ and g are new universal constants. Equation (15) leads to forces with a range $1/\lambda$, and in order to account for the range actually observed we have to choose

$$1/\lambda = 3 \times 10^{-13} \text{ cm.} \qquad \text{.....(16)}$$

The constant g corresponds to the electric charge of a particle in the electromagnetic case. In our case g is regarded as a new type of charge which determines the strength of the new nuclear field ϕ surrounding a neutron or proton. Both these particles are assumed to have the same g charge. g can be determined from the binding energy of the deuteron. At a distance $1/\lambda$ a proton and neutron have a potential energy of about 25×10^6 ev. Thus $g^2\lambda \simeq 25 \times 10^6$ ev. or

$$g \sim 5 \text{ to } 7 \text{ electron charges.} \qquad \text{.....(17)}$$

So far the new theory is a classical theory. In the same way as the electromagnetic field, the new field has to be quantized. The quantization of the electromagnetic field leads to the existence of light quanta and in a similar way we shall obtain a new type of quanta arising from the quantized nuclear field ϕ. While the light quanta arising from the electromagnetic field satisfy a wave equation of the type

$$\nabla^2\phi - \frac{1}{c^2}\frac{\partial^2\phi}{\partial t^2} = 0,$$

the wave equation for the Yukawa-potential will now be of the type

$$\nabla^2\phi + \lambda^2\phi - \frac{1}{c^2}\frac{\partial^2\phi}{\partial t^2} = 0. \qquad \text{.....(18)}$$

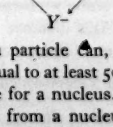
Equation (18) can be regarded as a wave equation for a particle with a rest mass

$$m = \frac{\lambda\hbar}{c} \simeq 150 \text{ electron masses.}$$

There is another very important feature of the new field ϕ. From the saturation properties of the nuclear forces we know that the forces between a neutron and a proton are mainly exchange forces, i.e. they are connected with a steady permanent exchange of charge between the two particles. The proton gives its positive charge to the neutron becoming itself a neutron whilst the neutron turns into a proton. Forces of this type are very familiar in quantum physics, they are, for instance, responsible for the binding in a homoeopolar molecule. In order now to account for this exchange nature of the nuclear forces one has to attribute an ordinary electric charge to the new field ϕ. The quanta arising from this quantized field will therefore also be charged, and it can be seen that they can have a positive or negative elementary charge e. Let us denote these charged quanta with a rest mass $m \simeq 150$ electron masses by Y^+, Y^- respectively. The production of the new field ϕ by a proton or neutron can then also be considered as a virtual emission of such a Y particle by a proton or neutron as indicated by the equations (just as the electromagnetic field of an electron can be considered as a virtual emission of light quanta)

$$\left.\begin{aligned} N &\rightleftarrows P + Y^-, \\ P &\rightleftarrows N + Y^+. \end{aligned}\right\} \qquad \text{.....(19)}$$

The charge is, of course, conserved in these processes. The forces between a proton and a neutron can then also be considered as due to a mutual emission and reabsorption of those particles,



An actual emission of such a particle can, of course, only take place if the necessary energy—mc^2 must be equal to at least 50 or 100 times 10^6 ev.—is available, and this is in general not the case for a nucleus. Only cosmic rays have sufficient energy to liberate those particles from a nucleus. If, for instance, a neutron or light quantum with an energy of at least 100×10^6 ev. passes through a nucleus it might be expected that such a quantum would be emitted.

From the equations (19) we can deduce another important property of the quanta $Y^\pm$. Since both the proton and neutron have a spin $\frac{1}{2}$ and obey the Pauli exclusion principle, it follows from general laws of quantum theory that the particles Y^+, Y^- have an integral spin and obey Bose statistics just as is the case for light quanta but in contrast to electrons, protons and neutrons.

The quanta Y^+, Y^- have exactly the properties which the heavy electrons occurring in cosmic radiation are now found to have. The mass calculated from the range of the nuclear forces is in excellent agreement with that measured on cosmic-ray tracks, table 2. It was in fact by the above arguments that Yukawa predicted the existence of heavy electrons three years before they were actually discovered.

There can be no doubt that the ideas developed above are in principle correct and that the heavy electrons found in cosmic rays are the quanta responsible for the nuclear field. However, in the form in which the new theory was originally published by Yukawa, it was not in agreement with all the experimental facts. It turned out, for instance, that the proton-neutron potential derived from the theory leads to a ^{1}S state for the lowest state of the deuteron whilst the ^{3}S state turned out to be repulsive. The experiments show that the ground state is a ^{3}S state and that the ^{1}S state, though still attractive, has a potential of only half of that of the ^{3}S state. Thus the theory had to be modified in some way.

Nuclear forces and magnetic moments. After the existence of the heavy electrons was made probable by cosmic-ray experiments a consistent theory based on the general ideas discussed above has been developed independently by Fröhlich, Heitler and Kemmer[32,44], by Yukawa, Sakata and Taketani[67] and by Bhabha[11a]. The main point of the new theory is the following: The nuclear field is not described by a single potential ϕ as in Yukawa's original paper but in a similar way to the electromagnetic field, by four potentials, a scalar potential ϕ and a vector potential $\mathbf{A}$, from which six components of field strengths can be derived. The field equations are almost identical with Maxwell's equations, the only difference being that terms also occur which contain the universal length λ or the rest mass m. All four potentials satisfy in vacuo the wave equations (18) and an equation corresponding to the Lorentz condition

$$\operatorname{div}\mathbf{A}+\frac{1}{c}\frac{\partial\phi}{\partial t}=0.$$

Apart from the terms $\lambda^2\phi$, $\lambda^2\mathbf{A}$ these equations are identical with Maxwell's equations. It turns out that heavy electrons can exist in different states of polarization, but, in contrast to light quanta, quanta with longitudinal polarization also exist.* This can also be expressed in the following way: In the same sense as a light quantum, the heavy electron has a spin $1\hbar$.

In addition to the field equations there is an interaction of the heavy electron with the nuclear particles of such a form that the heavy electron can be emitted and absorbed by a proton and neutron, equation (19).†

There is another point in which Yukawa's original theory had to be extended in order to account for all the important experimental facts. From the scattering of two protons[62,21] it is known that two protons have also an attractive potential of the same order of magnitude as a proton and neutron. The proton-neutron field

* The existence of the latter is a consequence of the finite rest mass.

† This interaction is found to depend upon the spin direction of the proton (neutron) and contains two universal constants g, f, both with the dimensions of a charge and of the order of 5 electron charges. They relate respectively to the longitudinal and transverse heavy electrons.

cannot, of course, be of an exchange type because no exchange of charge can take place between two particles with the same charge. The quantization of the proton-proton field leads, therefore, to quanta which carry no charge but have the same rest mass m. We call these neutral particles "neutrettos". The theory has, therefore, to be built up in such a way that it includes also neutral particles. This can be done in an exceedin glysimple way as is shown by Kemmer[45]. We shall not discuss the details of this theory here but confine ourselves to a discussion of the main results.

The main results of this theory are the following:

(*a*) The lowest state of the deuteron is a ^3S state which is always attractive.

(*b*) The ^1S state of the deuteron is higher than the ^3S state and the constants occurring in the theory can be chosen so that its position is in agreement with experiment.

(*c*) The proton-proton potential is exactly equal to the proton-neutron potential in the ^1S state.

(*d*) Taking the mass m of the heavy electron from cosmic rays all forces have the range $1/\lambda \simeq 3 \times 10^{-13}$ cm. in agreement with experiment. For S states the proton-neutron and proton-proton potentials have the form const. $\times e^{-\lambda r}/r$.

(*e*) The theory leads to an anomalous magnetic moment of the proton μ_P and neutron μ_N which are of the order of magnitude actually observed and have the right sign (positive for the proton, negative for the neutron).

(*f*) The ratio of the anomalous magnetic moments• depends only upon the mass of the heavy electron (and of the proton) and can therefore be calculated exactly. It is in good agreement with the experiments[31].†

The existence of an anomalous magnetic moment can be understood in the following way. A free proton is surrounded by a nuclear field or, in other words, is capable of emitting a positive heavy electron transforming itself into a neutron. A free proton can, of course, not actually emit a heavy electron (because the energy is not available) but it will spend a certain fraction of time in a neutron state in which it has virtually emitted a heavy electron. Consider now a free proton with its spin directed upwards. Then the theory shows that it will spend a certain fraction of time as a neutron and a heavy electron with an angular momentum $+1\hbar$ about the axis of spin direction. Because of the conservation of angular momentum the neutron has then a spin directed downwards. In this state the heavy electron will have a large magnetic moment—about 12 times larger than a Bohr nuclear magneton—because its mass is about 12 times smaller than that of the proton. The fraction of time which the proton spends in this dissociated state is calculable from the

• Subtracting from the magnetic moment of the proton the normal Dirac-magneton which is equal to one Bohr nuclear magneton μ_0.

† The theory gives the formula

$$\frac{\mu_P - \mu_0}{-\mu_N} = \frac{M}{M+m} - \frac{m}{M},$$

where M is the mass of the proton and m the mass of the heavy electron.

theory and is of the order $g^2/\hbar c$. The additional magnetic moment of the proton is, therefore, in units of the Bohr nuclear magneton μ_0

$$\frac{\mu_P - \mu_0}{\mu_0} \simeq \frac{g^2}{\hbar c}\frac{M}{m} \simeq 2, \qquad \text{.....(20)}$$

which is just of the right order of magnitude. The same consideration applies to the neutron, the sign of μ_N is the negative one because a neutron emits a negative heavy electron.

Although there can be little doubt that this theory is correct in its outline, it must, of course, only be regarded as a preliminary theory. In particular it has turned out that it cannot be applied for very fast heavy electrons with velocity nearly equal to that of light.* This point will be important for the application to cosmic rays.

Heavy electron and cosmic rays. Very little can be said yet about what happens when a heavy electron passes through matter. The mass of the heavy electron is so large that the energy loss by radiation (which is proportional to $1/m^2$) is negligible. On the other hand, the measurements of the energy loss (figure 7, § 6) indicate that the energy loss is still very much larger than that by ionization only. There must be, therefore, some other processes by which a heavy electron can lose energy. The experiments discussed in § 2 show that a heavy electron is also capable of producing showers and bursts. If we consider the part of the Rossi curve, figure 4, which is due to showers produced by the penetrating rays, it seems that there are two types of showers which the latter can produce. The curve (b) rises within the first few cm. of lead very steeply and remains then practically constant. This indicates a type of shower consisting of particles with a range of only a few cm. of lead. These particles are certainly electrons with an energy of not much more than 10^8 ev. Thus we expect a heavy electron to be capable of producing showers consisting of electrons of comparatively small energy. At a thickness of 15 to 20 cm. of lead the shower curve rises again and has a second maximum. This shows that some of the shower particles have themselves a penetrating power up to 20 cm. of lead and can, therefore, only be heavy electrons themselves. Thus a heavy electron is also capable of creating showers consisting of heavy electrons.

We might also mention that several authors have come to the conclusion that a part of all these showers is due to a non-ionizing (i.e. neutral) radiation with a penetrating power similar to, or even larger than, that of the heavy electrons[7,48].

It seems that the heavy electron is far from being a stable particle and disintegrates in some way as soon as it reaches an energy less than 2×10^8 ev., because practically no heavy electrons are observed below this energy.† This would lead to the conclusion that the heavy electrons cannot themselves come from outside the earth (their life-time would not be long enough to travel astronomical distances),

* For the reasons for the breakdown of the theory at high energies the reader is referred to the original papers[32,44,67].

† It has been suggested by Yukawa[65] and also by Bhabha[11] that the heavy electron can disintegrate spontaneously, without the presence of other particles, into an ordinary electron and neutrino (β decay of the heavy electron).

but are created as secondaries by the primary particles, e.g. by electrons, in the atmosphere. The same conclusion has also been reached by several authors[20] from indirect experimental evidence concerning the energy balance in the atmosphere. Although no direct proof of this fact is available yet, we think it very probable that the heavy electrons are secondaries and that the electrons with the primary energy spectrum described in § 5 are the only really primary particles coming from the universe.

We shall not discuss in this report all the different opinions and views which have been expressed about the behaviour of the heavy electrons in cosmic radiation so far; they will certainly be subjected to considerable alterations in the near future.* But one question re-presents itself in connexion with the theory of the heavy electron, namely: is the theory, which has been successful in the explanation of the nuclear properties, capable of accounting for the behaviour of the heavy electrons in cosmic radiation, and what are its consequences?

Unfortunately, the theory developed in the earlier part of this section is only valid for small energies and breaks down for cosmic-ray energies. In spite of this fact, it might be hoped that the theory would give at least some qualitative information about the processes to be expected, and since it exhibits some features which are not common to any other of the existing theories, we shall briefly mention some of its consequences.

The main assumption made in this theory is expressed in the equations (19) stating that in suitable circumstances a heavy electron can be absorbed or emitted by a nuclear particle. If such an absorption takes place the energy of the primary heavy electron must be given up in some way. This leads to a number of different processes which could be expected to occur if a heavy electron passed through a nucleus[40]:

(*a*) A heavy electron, Y^- say, is absorbed by a proton, and re-emitted into another direction. This process represents simply an anomalous scattering by a nucleus. A certain amount of energy is given up to the nucleus (in order to conserve momentum) and the primary electron loses a certain amount of energy. Possibly this or similar processes are responsible for the additional energy loss of heavy electrons.

(*b*) A heavy electron is absorbed by a nuclear particle, the energy being given up in the form of a light quantum

$$Y^- + P \rightarrow N + h\nu. \qquad \text{.....(21)}$$

Subsequently, the light quantum will create an ordinary cascade shower. Thus an electron shower is produced, as was in fact seen to be the case from the shape of the Rossi curve.

Electron showers can also be produced if a heavy electron hits an electron in an atomic shell in a direct head-on collision, transferring a large amount of energy to the latter. The fast electron will then produce a cascade shower. The latter type of process leads, however, only to comparatively small showers[10].

* Half-empirical attempts to bring the observations into a consistent scheme have been made by Nordheim[33] and Euler[30].

(*c*) A very surprising and perhaps the most important feature of the theory of the heavy electron is the following. If a heavy electron of sufficiently high energy, e.g. larger than about 3 to 5 times 10^8 ev., interacts with a nucleus, a so-called multiple process or explosion may occur. It may then be that a large number of heavy electrons is emitted simultaneously in one elementary act, for instance:

$$Y^- + P \rightarrow P + Y^- + Y^+ + Y^-. \qquad \text{.....(22)}$$

The number of heavy electrons emitted is only restricted by the initial energy of Y^-. The existence of those explosive processes was first postulated by Heisenberg[38] on the grounds of Fermi's theory of β decay and made responsible for the majority of big showers and bursts. It has since been shown by Heitler[40] that the theory of the heavy electron leads to the same qualitative consequence and we believe now that it is the type of process given by equation (22) which actually is responsible for the explosions occurring in nature.

The process (22) gives rise to the production of a shower consisting of heavy electrons, i.e. of penetrating particles only. Penetrating showers really occur in nature, we have seen that they are responsible for the second maximum of the Rossi curve. In particular a discussion of the transition curve for bursts by Euler[30] has shown that a large fraction of the bursts must be ascribed to multiple processes of the type indicated by equation (22).

(*d*) So far we have only considered processes for which the nuclear particles can be considered as free. In a heavy nucleus this is not quite the case. In all the processes (*a*), (*b*), (*c*) a certain amount of energy is transferred to the nuclear particle. According to Bohr's theory the latter will interact at once with other nuclear particles which results in a general excitation of the whole nucleus. It might then happen that the nucleus becomes very highly (10^8 ev. and more) excited. In course of time the nucleus will release a number of its constituents, emitting subsequently a few protons and neutrons in all directions until its energy has decreased to the ground state. Thus, we shall find a shower containing heavy particles. Those nuclear explosions have occasionally been observed, we have discussed them in § 2. They can, of course, also be produced by a direct head-on collision of any very energetic cosmic-ray particle with a nucleus[36].

(*e*) Finally, we might mention that we should also expect neutral penetrating particles in cosmic rays. From the discussion of the proton-proton forces it has been seen to be probable that these neutral particles with mass $m \simeq 150$ electron masses also exist. We have called them "neutrettos". In cosmic rays they give rise to almost the same processes as the charged heavy electrons—except of course those for which the charge is essential. In particular, it can happen that a heavy electron is transformed into such a neutretto (denoted by Y^0) during a collision with a nuclear particle, for instance:

$$Y^- + P \rightarrow N + Y^0, \qquad \text{.....(23)}$$

and vice versa. In this way neutrettos can be emitted and absorbed. They give rise also to all the shower processes of the types (*c*) and (*d*), but not of the type (*b*) because for the emission of light the charge is essential. It seems that there are

some indications in cosmic rays for those showers produced by neutral particles[4], but it would lead too far to discuss them in detail here.

(*f*) On the basis of this theory it can also be understood that the heavy electrons are secondaries created by ordinary electrons and light quanta. The reverse process of (*b*) for instance, gives rise to a creation of a heavy electron by a light quantum in a collision with a nucleus, and a rough estimate shows that the cross-section is large enough to account for the actual numbers of heavy electrons created in the high atmosphere.

On the whole our knowledge about the hard cosmic radiation is still very poor, both from the experimental and theoretical point of view. On the other hand, the situation has been clarified sufficiently to reveal the intimate connexion between these particles and the nuclear properties. It is clear that future experiments on the hard cosmic rays could give substantial contributions to a further deepening of our knowledge about the fundamental properties of the elementary particles.

REFERENCES

This list does not aim at completeness and contains only a small proportion of the papers connected with the subjects dealt with in this report.

(1) Anderson. *Phys. Rev.* **44**, 406 (1933).
(2) Anderson and Neddermeyer. *Phys. Rev.* **50**, 263 (1936).
(3) Arley. *Proc. Roy. Soc.* A, **168**, 519 (1938).
(4) Arley and Heitler. *Nature, Lond.*, **142**, 158 (1938).
(5) Auger, Ehrenfest and Leprince-Ringuet. *J. Phys. Radium*, **7**, 58 (1936).
(6) Auger and Grivet-Meyer. *C.R. Acad. Sci., Paris*, **204**, 572 (1937).
(7) Barnóthy and Forró. *Z. Phys.* **104**, 744 (1937).
(8) Bethe and Heitler. *Proc. Roy. Soc.* A, **146**, 83 (1934).
(9) Bhabha and Heitler. *Proc. Roy. Soc.* A, **159**, 432 (1937).
(10) Bhabha. *Proc. Roy. Soc.* A, **164**, 257 (1938).
(11) Bhabha. *Nature, Lond.*, **141**, 117 (1938).
(11*a*) Bhabha. *Proc. Roy. Soc.* A, **166**, 501 (1938).
(12) Blackett and Occhialini. *Proc. Roy. Soc.* A, **154**, 573 (1936).
(13) Blackett and Brode. *Proc. Roy. Soc.* A, **139**, 699 (1933).
(14) Blackett. *Proc. Roy. Soc.* A, **159**, 1 (1937).
(15) Blackett and Wilson. *Proc. Roy. Soc.* A, **165**, 209 (1938).
(16) Blackett. *Proc. Roy. Soc.* A, **165**, 11 (1938).
(17) Blau and Wambacher. *S.B. Akad. Wiss. Wien*, IIa, **146**, 623 (1937).
(18) Bøggild. "Über die sekundären Wirkungen der Höhenstrahlen." Dissertation (Copenhagen, 1937).
(19) Bowen, Millikan and Neher. *Phys. Rev.* **52**, 80 (1937).
(20) Bowen, Millikan and Neher. *Phys. Rev.* **53**, 217 (1938).
(21) Breit, Condon and Present. *Phys. Rev.* **50**, 825 (1936).
(22) Brode and Starr. *Phys. Rev.* **53**, 3 (1938).
(23) Carlson and Oppenheimer. *Phys. Rev.* **51**, 220 (1937).
(24) Carmichael. *Proc. Roy. Soc.* A, **154**, 223 (1936).
(25) Clay. *Proc. Acad. Sci. Amst.* **35**, 1282 (1932).
(26) Compton and Stephenson. *Phys. Rev.* **45**, 441 (1934).
(27) Ehmert. *Z. Phys.* **106**, 751 (1937).
(28) Ehrenfest. *C.R. Acad. Sci., Paris*, **206**, 428 (1938).
(29) Epstein. *Phys. Rev.* **53**, 862 (1938).
(30) Euler. *Naturwissenschaften*, **26**, 382 (1938).
(31) Fröhlich and Heitler. *Nature, Lond.*, **141**, 37 (1938).

(32) Fröhlich, Heitler and Kemmer. *Proc. Roy. Soc.* A, **166**, 154 (1938).
(33) Fussel. *Phys. Rev.* **51**, 1005 (1938).
(34) Geiger. *Ergebn exakt. Naturw.* (1935).
(35) Heisenberg. *Z. Phys.* **101**, 533 (1936).
(36) Heisenberg. *Ber. sächs. Ges. (Akad.) Wien*, **89**, 369 (1937).
(37) Heitler. *Quantum Theory of Radiation* (Oxford, 1936).
(38) Heitler. *Proc. Roy. Soc.* A, **161**, 261 (1937).
(39) Heitler. *Nature, Lond.*, **140**, 235 (1937).
(40) Heitler. *Proc. Roy. Soc.* **166**, 529 (1938).
(41) Grivet-Meyer. *C.R. Acad. Sci., Paris*, **206**, 833 (1938).
(42) Hoffmann. *Int. Conf. Physics, London* (1934).
(43) Janossy. *Z. Phys.* **104**, 430 (1937).
(44) Kemmer. *Proc. Roy. Soc.* A, **166**, 127 (1938).
(45) Kemmer. *Proc. Camb. Phil. Soc.* **34**, 354 (1938).
(46) Lemaître and Vallarta. *Phys. Rev.* **43**, 91 (1933).
(47) Leprince-Ringuet and Crussard. *J. Phys. Radium*, **8**, 207 and 213 (1937).
(48) Maass. *Ann. Phys., Lpz.*, **27**, 507 (1936).
(49) Montgomery and Montgomery. *Phys. Rev.* **53**, 955 (1938).
(50) Neddermeyer and Anderson. *Phys. Rev.* **51**, 884 (1937).
(51) Nishina, Takeuchi and Ichimiya. *Phys. Rev.* **52**, 1198 (1937).
(52) Nordheim. *Phys. Rev.* **51**, 1110 (1937).
(53) Nordheim. *Phys. Rev.* **53**, 694 (1938).
(54) Pfotzer. *Z. Phys.* **102**, 23 and 41 (1936).
(55) Regener and Pfotzer. *Nature, Lond.*, **136**, 718 (1935).
(56) Rossi. *Int. Conf. Physics, London* (1934).
(57) Ruhlig and Crane. *Phys. Rev.* **53**, 266 (1938).
(58) Schmeiser and Rothe. *Ann. Phys., Lpz.*, **32**, 161 (1938).
(59) Schwegler. *Z. Phys.* **96**, 62 (1935).
(60) Street and Stevenson. *Phys. Rev.* **52**, 1003 (1937).
(61) Trumpy. *Nature, Lond.*, **141**, 909 (1938).
(62) Tuve, Heydenburg and Hafstadt. *Phys. Rev.* **50**, 806 (1936).
(63) Williams and Pickup. *Nature, Lond.*, **141**, 684 (1938).
(64) Wilson. *Proc. Roy. Soc.* A, **166**, 482 (1938).
(65) Yukawa. *Proc. Phys. Math. Soc. Japan*, **17**, 48 (1935).
(66) Yukawa and Sakata. *Proc. Phys.-Math. Soc. Japan*, **19**, 1084 (1937).
(67) Yukawa, Sakata and Taketani. *Proc. Phys.-Math. Soc. Japan*, **20**, 319 (1938).

THE MECHANISM OF THE GEIGER COUNTER

By A. NUNN MAY, Ph.D.

Wheatstone Laboratory, King's College, London

§ 1. INTRODUCTION

The first method of detecting single particles by their ionization was that developed thirty years ago by Rutherford and Geiger[5]. They used the principle of ionization by collision which had then recently been discovered by Townsend[6]. Their apparatus was very similar in principle to the modern form of counter. A fine wire stretched down the axis of a brass tube was insulated from it at the two ends by ebonite stoppers. The pressure of air in the tube was reduced to a few centimetres of mercury and a potential near the break-down potential of the gas was applied to the tube, the outer case being negative. The wire was connected to a quadrant electrometer.

Under these conditions any electrons set free by ionization in the gas were dragged to the wire and in the strong electric field in its neighbourhood gave rise to fresh positive ions and electrons by collision with the molecules of the gas. These electrons ionized the gas by collision in their turn, and in this way an electron avalanche was set up in which each primary electron was multiplied by a factor which was of the order of 1000. With this multiplication the ionization from a single α-particle was easily detected by the quadrant electrometer.

In subsequent experiments in order to reduce the number of natural particles this apparatus was replaced by a ball counter in which the multiplication took place at the surface of a small sphere. As we shall see later these methods of proportional multiplication with changes in detail are still often used for the detection of heavily ionizing particles. They have however been largely replaced by the method introduced by Greinacher[7] in 1927 in which a parallel plate ionization chamber in combination with a sensitive amplifier is used. Wynn-Williams[2] has already given an account of this technique in these reports. It has the advantage of giving a more direct measure of the amount of ionization produced by the particle, since only the primary ionization is collected without any multiplication. This method can only be used when the primary ionization is sufficient to give a change in potential on the electrode larger than the background fluctuations of the amplifier. This makes it inapplicable to measurements on β-rays, and for them some system of multiplication by collision is generally used.

For many purposes the proportional multiplication of ionization is not essential, only the *number* of particles entering the counter being required and not their ionization. For these experiments because of its simplicity the non-proportional

counter is most suitable. The first of these was introduced by Geiger[8] in 1913. This consisted of a wire with a fine point at its end which was kept at a high (positive or negative) potential with respect to an outer case. Any ionization in the neighbourhood of the point led to a momentary current in which the charge flowing was independent of the primary ionization. This trigger action made the counter very useful for the detection of β-rays and indeed in any experiments involving the detection of small amounts of ionization. It found many applications and there is an extensive literature concerning its mechanism and methods of preparation. These will not be dealt with here since they will be found summarized in the German Handbuchs[1] and because this type of counter has been almost completely superseded by the tube-counter introduced by Geiger and Müller[9] in 1928.

This was very similar in form to the original proportional counter of Geiger and Rutherford but the potential across the tube was raised. In certain circumstances the counter then responded to ions in the same way as the point counter, by a single abrupt discharge. It had the great advantage of being sensitive over nearly its whole volume instead of just in the neighbourhood of a point. The reasons why the discharge when once started is only momentary and not continuous were and still are rather uncertain. Geiger and Müller held at first that a surface film of oxide on the wire was essential. Other investigators found that provided the surface of the wire was smooth its material did not matter provided the counter was in series with a large resistance. On the other hand the surface of the outer electrode and the purity of the gas were very important. These points will be discussed later.

Because of its much greater sensitive volume and reliability the Geiger-Müller counter immediately found many applications, and special forms were developed for work with x-rays, photometry, etc. This led to the need for elaborate auxiliary circuits for rapid counting, for detecting coincident discharges in several counters, and for maintaining constant potentials across the counter[2, 18, 31]. At the same time in order to obtain reliable counters the mechanism of the discharge was studied with careful purification of the gas and cleaning of the metal surfaces with the technique which had been developed for high vacuum work.

Several summaries of the literature on Geiger-Müller counters have recently been published among which one may mention an excellent article by Moon[3] and a very recent book by Thibaud, Cartan and Comparat[4]. To avoid covering the same ground as these this article will deal mainly with the problem of the mechanism of the counter. Owing to the work of Werner[22] and others it is now possible to give a fairly systematic account of this aspect of the subject.

§ 2. THE PROPORTIONAL COUNTER

Before considering the recent work on the non-proportional discharge we will discuss in some detail the proportional multiplication counter. This has often been used for detecting primary ionization which would be too small to be easily detected directly as in the counting of fast protons, or when the very rapid discharge and consequent short resolving time of the counter is an advantage as when experiments are being carried out in the presence of intense γ-radiation[10, 11, 12]. Quite apart from

these applications the processes in the multiplication counter are important for an understanding of the more complicated non-proportional discharge.

The simplest form of apparatus for experiments on multiplication by collision consists of two plane parallel electrodes at a variable distance apart. This is the arrangement which was used in the classical experiments of Townsend[6].

This arrangement has not often been adopted for counting purposes since a higher potential is needed at any given pressure than in the wire counter, and the multiplication is not the same for ions formed at different points in the counter, increasing with the distance from the anode.

However Zipprich[13] has used this method together with a wire grid arranged so that the ions were formed in a nearly field-free space from which they drifted into the multiplication region. It was then found that the charge collected was proportional to the number of primary ions and Townsend's equations for the variation of the multiplication factor were verified. By an ingenious system using two successive stages of multiplication Winkler[14] has extended this method so far as to detect single ions proportionally with a multiplication of 200,000 times.

The most commonly used proportional counters have been the ball and wire types. The first of these was studied in detail by Geiger and Klemperer[10] and the second by Geiger and Zahn[11]. In these the multiplication is confined to the surface of a wire or sphere of small diameter which forms the anode. Using the Townsend theory expressions can be derived for the variation of the multiplication factor with the diameters of the electrodes, pressure of the gas and the voltage. These have been verified experimentally by Müller[15] for air at small multiplication factors. In designing such a counter the working voltage can be deduced from these formulae or more easily from those given by Werner[22] for the non-proportional counting voltage. The potential in the proportional region will not be far short of this counting voltage.

It is important that the multiplication should be independent of the point in the counter at which primary ions are formed and should not be subject to large random fluctuations. In the wire and ball counters the multiplication is confined to a small volume so that ions formed outside this region should all be multiplied to the same extent. Nevertheless multiplication by collision is an essentially random process and must involve statistical fluctuations, which would not be revealed in the classical experiments on steady ionization currents. This point has not been very much studied experimentally. Geiger[16] gives a distribution curve for the size of kick obtained with α-particles entering a wire counter in which the fluctuations appear to be small, of the order of 1/40 of the mean, but Brubaker and Pollard[17] have found fluctuations of the order of 10 per cent in various gases and also find a decrease in the multiplication for ions formed at a distance from the wire. They ascribe this last to the capture of electrons to form negative ions.

There have not been any theoretical papers on this aspect of multiplication. Some calculations by the writer indicate that the fluctuation in the final amount of ionization should be of the order of $M\sqrt{N}$, where N is the number of primary ions, and M the multiplication factor. It would be interesting to compare this prediction with experiments such as the above but the data given are not sufficient.

In the writer's experience the multiplication counter has been very useful, and simplifies the counting of particles very considerably by making it possible to use linear amplifiers of very much less sensitivity than in the direct method. Their only disadvantage is that the multiplication varies rapidly with the voltage applied, so that this must be constant within narrow limits (< 1 v. at 1000 v.). Circuits ensuring this have been developed[18], and by frequent calibration with standard sources of ionization the counters can be made quite reliable.

A different kind of proportional counting which has recently come into prominence is that using the electron multiplier devised by Zworykin[19]. This depends on the fact that a fast electron colliding with a surface can cause the emission of a large number of secondary electrons. These can be accelerated and directed on to a similar surface to give a still greater number. By arranging several such stages in series, very high multiplication factors can be attained. Z. Bay[20] and W. H. Rann[21] have shown that this device can be used to detect β-particles which are directed on to the first surface of the multiplier. The great advantage claimed for this type of counter is the small resolving time. The multiplication process takes about 10^{-8} sec. which is much smaller than can be attained with any gas-filled counter.

§ 3. THE NON-PROPORTIONAL COUNTER*

If the potential across a wire counter is raised so that the multiplication factor becomes very large the tube may break down in one of several ways when a multiplication factor of the order 10^4 is reached. It may give a continuous or intermittent discharge which once started does not depend on the presence of any external source of ionization, or it may *count*, i.e. give a single abrupt discharge for each particle entering the tube. There are a number of different ways of securing this behaviour of which the simplest is to place in series with the tube a high resistance, R, of the order of 10^8 to 10^9 Ω. The circuit is then as shown in figure 1.

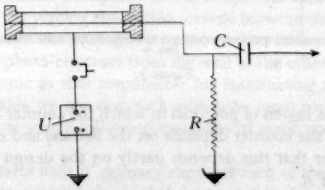


Figure 1.

The discharges are detected by means of the momentary current through R. A small condenser C feeds the potential change due to this current into the recording circuit.

* The following account is mainly founded upon the papers of Werner[22] and Cosyns[23] to which reference should be made for further details.

If with this arrangement and with a constant source of ionization the number of discharges per min. is determined for various values of the applied voltage U, a characteristic curve is obtained, such as figure 2.

This shows a sharp rise from the starting potential V_0 (at which the potential changes on the wire are too small to be recorded) to a constant value giving a plateau BC. This plateau extends to a certain over-voltage $U - V_0$ where the number of discharges begins to rise sharply, along the curve CD.

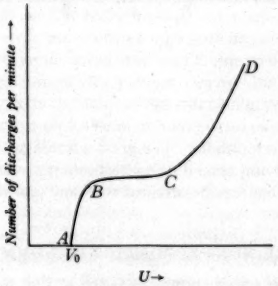


Figure 2.

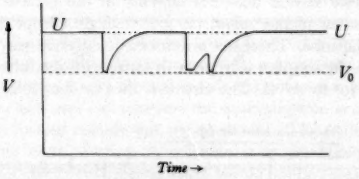


Figure 3.

This plateau is the region of potential in which the counter is normally operated and the reliability of the counter depends on the flatness and extent of this region. It will be shown later that this depends partly on the design of the counter, and partly on the value of R.

A more detailed study of the discharge may be made by means of the cathode ray oscillograph,* the potential V across the counter being applied through a suitable circuit to the Y plates, while a time base circuit is applied to the X plates so that the spot on the fluorescent screen gives a plot of the variation of V with time[22, 24, 25, 26]. The type of result obtained is shown in figure 3.

* Similar results can of course be obtained with any form of rapid oscillograph.

This may be accounted for by the following sequence of events. The potential difference V across the counter is initially U. The onset of the discharge causes the internal resistance of the counter to fall from infinity to a fairly low value (about 1 MΩ.), and the potential V therefore drops sharply. This fall is arrested when V reaches a value V' slightly above the starting potential V_0. The discharge may then burn steadily for a very short time, or stop instantly. As soon as the discharge ceases the potential across the counter returns exponentially to its original value U with a time constant given by the total capacity across the counter, C', and the resistance R. The potential change communicated to the recording system is thus $U-V'$ which is nearly equal to the over-voltage.

The time occupied by a kick is mainly determined by the recovery time, and is therefore of the order RC'. If another discharge occurs before the first has quite disappeared, the potential change will be less than the over-voltage, as shown in figure 3, and such a discharge may fail to be recorded. It is therefore important for accuracy in rapid counting that the recovery time of the counting circuit should be as short as possible.

§ 4. THEORY OF THE COUNTER DISCHARGE

The proportional discharge consists of only one electron avalanche. If a continuous discharge is to be maintained fresh electrons must be supplied continually from outside the immediate neighbourhood of the wire. Many different mechanisms have been suggested for this: ionization of the gas by the positive ions, ejection of electrons by the positive ions when they reach the outer electrode, and the emission of photo-electrons from the outer electrode by photons from atoms excited in the electron avalanche. The relative importance of these processes in gas discharges in general is discussed in an article by Loeb[27] to which reference should be made. As far as Geiger counters are concerned the evidence is fairly conclusive that the last is the main process.* In Greiner's[28] experiments two counters A and B were used inside the same enclosure. It was found that a discharge in one always spreads to the other. By placing absorption screens between the two counters it was shown that the discharge was propagated by ultra-violet radiation which originating in one counter ejects photo-electrons from the wall of the other. This mechanism is most probably the same as that responsible for maintaining the discharge in one counter. Accepting this we see that each avalanche must produce enough photo-electrons in this way to give a following avalanche of equal size, if the discharge is to be continuous.

If an avalanche starts with N primary electrons each of these is multiplied by a factor m in the discharge so that the final charge arriving at the wire is nm electrons. The number of photo-electrons ejected from the outer wall will be proportional to Nm, say pNm. For the discharge to be continuous we must have

$$m=\frac{1}{p}=M, \qquad \text{.....(1)}$$

* Cosyns[23] and Thibaud *et al.*[4] following him assume that electrons are provided by the first process above, ionization by positive ions, but the weight of evidence seems to be against this.

where M, which we shall call the normal multiplication, is generally of the order of 10^5. p and therefore M will depend on a number of factors such as the pressure and the work function of the surface of the outer electrode. In particular it will be proportional to the solid angle subtended by the outer electrode at the wire. Christoph and Hanle[20] have shown that if this angle is reduced the counting voltage is increased, as we should expect.

The normal multiplication is attained when the field near the wire reaches a certain value, and this will occur at the starting potential V_0. Werner[22] has calculated the starting potential on the assumption that the potential fall in n electron free paths of length K/p next the wire has to exceed a critical value u. This condition can easily be calculated and gives the following formula for the starting potential:

$$V_0 = \frac{u \log_e r_a/r_i}{\log_e (Kn/pr_i + 1)}. \qquad \ldots\ldots(2)$$

r_a and r_i are here the radii of the outer and inner electrodes.

This formula has been tested by Werner and by Haines[30] for a number of different gases. It gives the variation of V_0 with p and with r_i very well. The values of the constants u and n are determined empirically.

At the starting potential the initial multiplication is equal to the normal. As soon as the discharge has started, however, the positive ions formed near the wire by successive avalanches accumulate, and their space charge modifies the field. Werner has calculated this effect. In order to maintain the same field at the surface of the wire the voltage across the tube has to be raised by an amount $V - V_0$ which to a first approximation is proportional to the density of the space charge and therefore to the positive ion current.

Thus
$$V - V_0 = ki, \qquad \ldots\ldots(3)$$

where k is an internal resistance of the order of 1 MΩ. The value of k can be calculated in terms of the other constants.

If the counter is used in the circuit of figure 1 we can calculate the current in the discharge for a given over-voltage. From equation (3) we shall have

$$i = \frac{U - V_0}{R + k}. \qquad \ldots\ldots(4)$$

This brings us to one of the main difficulties in explaining these discharges. Equation (4) is an equilibrium equation and the current i should be perfectly stable. But in fact as we have seen the current in the circuit stops when $U - V_0$ falls below a certain value. The use of the tube as a counter depends on this fact. Werner has investigated this point, and found that the discharge is unstable as long as i is less than a critical value $i_{\min}$ of the order of 1 μA. Above $i_{\min}$ a continuous discharge sets in. From equation (4) we can deduce that the over-voltage must not be allowed to become greater than

$$U_m - V_0 = (R + k)\, i_{\min}. \qquad \ldots\ldots(5)$$

This gives us an upper limit to the working range of the counter: the flat portion of the characteristic is generally considerably smaller than this. We see that in order

to obtain a long counting range it is necessary to use a large resistance R, which as we saw above increases the time occupied by each kick, and makes fast counting inaccurate.

This dilemma has recently been overcome by the use of special value circuits. These will not be dealt with here, they will be found summarized in a recent article by Johnson[31]. Briefly the principle of their action is to amplify the kick Ri on the wire by a factor μ and feed it back on to the counter. The condition (5) then becomes

$$U_n - V_0 = \mu\,(Ri_{\min}), \qquad \ldots\ldots(6)$$

so that for the same over-voltage the time of recovery can be μ times as short.

§ 5. INTERMITTENT NATURE OF THE DISCHARGE

Various theories have been put forward to account for the momentary nature of the discharge. The similar discharges in point counters were explained[32] in a similar way to the flashing neon tube discharges by the form of the current-voltage characteristic curve. If this is negative, i.e. if the voltage decreases with increasing current, then the discharge can be intermittent. Similar explanations have been offered[33] for the discharges in wire counters but Werner finds both theoretically and experimentally that except at very low pressures k is always positive, so that the discharge should be stable.

A very similar theory based on the action of space charge has been put forward by von Hippel[34]. This may be correct for the discharges between plane electrodes, but its extension to the wire counter seems to be unsound except in very abnormal circumstances[34a]. Here again Werner's work shows that space-charge effects lead only to the equilibrium condition, equation (4).

The presence of a surface layer on the cathode has been found by some authors to be essential. Very careful outgassing of the electrodes destroys the counting properties of the tubes in some cases[43,44]. This is probably because of the change produced in the work function of the surface. It seems unlikely that this surface layer actually extinguishes the discharge.

Werner himself suggests that the extinction occurs through random fluctuations in the positive ion space charge. This would occasionally lead to too low a field strength for the maintenance of the discharge. Werner does not give any quantitative estimate of this effect; a rough calculation shows that the fluctuations required would be too large to occur often enough. von Geel and Kerkum and, independently, the writer[35], have put forward the theory that the random fluctuations in the emission of the photo-electrons from the outer wall are responsible. For currents of the order of $i_{\min}$ the average number of photo-electrons generated by each avalanche is only of the order of ten, so that there is a finite chance that at some particular avalanche the number will be zero, when the discharge will stop. This theory gives a good agreement with Werner's data on the mean life time of the discharge at various values of the current i.

The value of i_{min} for a counter is a most important measure of its reliability. It will be seen from equation (5) that for a given value of R it determines the range of over-voltage in which the discharge is intermittent. It is therefore an advantage to have i_{min} as large as possible. Unfortunately at the moment none but empirical means can be used to predict its value. Werner gives tables of i_{min} for various gases from which it appears that it increases with pressure, and is large (1 μA.) for hydrogen, and very small for nitrogen and *pure* inert gases. It is to be hoped that the development of a correct theory of the intermittent discharge will enable values of i_{min} to be predicted, and these variations understood.

In the type of counter discussed above the discharge is extinguished by the current being brought below i_{min} by means of the resistance R. This type of discharge always gives a change in potential on the wire nearly equal to the over-voltage.

A quite different type of discharge has been obtained by Trost[36], in counters containing alcohol vapour mixed with the normal gas. These can be used with only a small series resistance, and a large over-voltage. The discharge is no longer extinguished by the external limitation of the current but by some internal mechanism. Trost found that the change in potential on the wire was not equal to the over-voltage, but changed if a condenser were placed in parallel with the counter in such a way as to show that a constant charge was being collected by the wire at each discharge.

Trost suggests that the mechanism of the counter is that the discharge lasts until a sufficient number of positive ions have been accumulated near the wire to bring the electric field throughout the counter below the value at which multiplication occurs. The discharge then ceases and this number of positive ions gives the charge collected by the wire: the magnitude of the charge actually collected agrees with this. It is not immediately clear however why the discharge should be extinguished instead of burning at some steady current i as the counters discussed above. The following suggestion may be offered. If the initial multiplication is very much greater than the normal multiplication, as it will be at high over-voltages, then the first few avalanches may form rather more than the number of positive ions required to reduce the multiplication to normal. The multiplication will then be less than normal until this excess charge has moved across the counter, and in this time the successive avalanches will have been decreasing and may have reached zero.

Because of the low series resistance with which they can be used these counters give a very quick resolving time (10^{-8} sec.) and a very uniform size of kick. Trost found that using argon with 2 per cent of alcohol vapour the counters could be operated at atmospheric pressure.

§6. SPURIOUS DISCHARGES AND OTHER DEFECTS

The plateau of the characteristic curve in figure 2 is hardly ever quite flat. The number of counts generally increased slowly with the voltage, an increase of 1 per cent per volt being quite usual. There are several causes for this. Any slight asymmetry of the electrodes leads to variations in the electric field in different parts of the counter, so that the critical value at which counting begins is not attained at the

same voltage over the whole tube. The effective counting volume therefore increases continually. Christoph[37] has investigated this effect by deliberately placing the wire at a variable distance from the axis of the outer cylinder. He found that even a small displacement had a marked effect on the plateau of the characteristic curve. On the other hand a local irregularity in the outer cylinder has little effect so that it is possible to use a wire gauze as the outer cylinder.

A similar effect is caused by the variation of the field at the ends of the outer cylinder. This can be avoided by shielding the wire in this region so that it does not count. In this way the counting volume is clearly defined, and has the same length at all voltages. Cosyns[23], by adopting this precaution, was able to obtain a very flat plateau.

Even with these precautions when the over-voltage reaches high values (100 v.) the number of discharges rises rapidly.

When U exceeds the value given by equation (5) the discharge becomes continuous, but well below that point, when the discharge is still intermittent, a large number of spurious discharges occur. These have been studied by various workers[38, 29, 23] whose results are rather complicated and contradictory. It seems to be fairly definite that many, if not all, of the extra discharges are induced discharges in the sense that some previous discharge has left the counter in an excited condition from which it recovers by a series of spurious discharges.

In Medicus' experiments[38] the distribution of the discharges in time was studied, and distribution curves plotted of the number of times the interval between successive discharges had certain values. For low voltages these agreed with the curves expected for a random distribution of discharges, but at high potentials there was a surplus of small intervals showing that one discharge had a tendency to cause another succeeding one. From these results Medicus found that the probability of an induced discharge decreased exponentially with the time after the parent discharge and this decay had a time constant of a few seconds. Other less detailed experiments such as those of Christoph and Hanle[29] and Cosyns[23] led to much the same results. Cosyns found induced discharges with a time delay of the order of minutes.

Medicus found that the presence of traces of water vapour in the counter encouraged these spurious discharges, a carefully outgassed tube did not show them. On the other hand many authors[43, 44] have found that counters with very carefully outgassed electrodes and purified gases fail to show any counting properties. This is cured by allowing a trace of some impurity, usually oxygen, to be present. It seems likely that this forms a surface layer on the cathode which is essential to the working of the counter. Lewis and Burcham[39] found that cleaning the outer tube with abrasives produced a number of spurious discharges which died away in about 10 min. These they ascribed to the emission of electrons in the process of oxidation of the cathode as the surface layer was reformed. It may be that a similar process causes the induced discharges discussed above. The heavy discharge at high voltages might clean part of the layer of the cathode, and during its reformation induced discharges would occur.

The effect of this surface layer is very strikingly shown in some of Trost's experiments in which the cathode surface was coated with a layer of varnish. This greatly extended the flat portion of the characteristic and made it possible to use cathodes of materials such as aluminium which gave very poor counters without it.

All counters, but especially the older types using ebonite insulators, are liable to erratic changes in the counting properties through the emission of gases from their walls. This may occur through a change in temperature or under the influence of the discharge. It can be avoided by connecting the counter to a large reservoir[40] which reduces the effect of a given amount of impurity, or better by using counters sealed in pyrex glass and carefully outgassed. The preparation of these has been described in many papers[23, 41, 42]. This is fairly simple if the counter is for detecting γ-radiation or cosmic rays, but when thin windows have to be used to admit β-rays the technical difficulties are greater but have been overcome.

As was mentioned above too thorough purification of the gas and cleaning of the electrodes destroys the counting properties of the tube. Many authors seem to be able to reach the right degree of impurity by using commercial gases without extra purification, or by accident. It is probably better to use deliberate mixtures of gases, such as those of the inert gases and hydrogen, or those recommended by Trost.

REFERENCES

General

(1) GEIGER. *Handb. Phys.* **22** (2), 155 (1933).
(2) WYNN-WILLIAMS. *Phys. Soc. Progr. Rep.* **3**, 239 (1936).
(3) MOON. *J. Sci. Instrum.* **14**, 189 (1937).
(4) THIBAUD, CARTAN and COMPARAT. *Quelques Techniques actuelles en Physique Nucléare* (Gauthier Villars, Paris 1938).

§1

(5) RUTHERFORD and GEIGER. *Proc. Roy. Soc.* A, **81**, 141 (1908); *Phil. Mag.* **24**, 618 (1912).
(6) TOWNSEND. *Electricity in Gases* (Oxford, 1914).
(7) GREINACHER. *Z. Phys.* **36**, 364 (1926); **44**, 319 (1927).
(8) GEIGER. *Phys. Z.* **14**, 1129 (1913).
(9) GEIGER and MÜLLER. *Phys. Z.* **29**, 839 (1928); **30**, 489 (1929).

§2

(10) GEIGER and KLEMPERER. *Z. Phys.* **49**, 753 (1928).
(11) GEIGER and ZAHN. See reference (1), p. 162.
(12) DUNCANSON and MILLER. *Proc. Roy. Soc.* A, **146**, 396 (1934).
(13) ZIPPRICH. *Z. Phys.* **85**, 592 (1933).
(14) WINKLER. *Phys. Z.* **37**, 557 (1936).
(15) MÜLLER. *Z. Phys.* **48**, 624 (1928).
(16) GEIGER. Reference (1), Fig. 7, p. 164.
(17) BRUBAKER and POLLARD. *Rev. Sci. Instrum.* **8**, 254 (1937).
(18) STREET and JOHNSON. *J. Franklin Inst.* **214**, 155 (1932). See also reference (4), p. 253.
(19) ZWORYKIN, MORTON and MALTER. *Proc. Inst. Radio. Engrs, N.Y.*, **24**, 351 (1936).
(20) BAY. *Nature, Lond.*, **141**, 284, 1011 (1938).
(21) RANN. *Nature, Lond.*, **141**, 410 (1938).

§3

(22) WERNER. *Z. Phys.* **90**, 384 (1934); **92**, 705 (1934).
(23) COSYNS and DE BRUYN. *Bull. Acad. Belg. Cl. Sci.* **20**, 371 (1934).
COSYNS. *Bull. tech. Ass. Ing. Brux.* p. 173 (1936).
(24) MEDICUS. *Z. Phys.* **74**, 350 (1932).
(25) RAMSEY and LIPMAN. *Rev. Sci. Instrum.* **6**, 121 (1935).
DANFORTH. *Phys. Rev.* **46**, 1026 (1934).
(26) SEARS and READ. *Phys. Rev.* **47**, 197 (1935).

§4

(27) LOEB. *Rev. Mod. Phys.* **8**, 267 (1936).
(28) GREINER. *Z. Phys.* **81**, 543 (1933).
(29) CHRISTOPH and HANLE. *Phys. Z.* **34**, 641 (1933).
(30) HAINES. *Rev. Sci. Instrum.* **7**, 411 (1936).
(31) JOHNSON. *Rev. Sci. Instrum.* **9**, 218 (1938).

§5

(32) APPLETON, EMELÉUS and BARNETT. *Proc. Camb. Phil. Soc.* **22**, 434 (1924).
TAYLOR. *Proc. Camb. Phil. Soc.* **24**, 251 (1928).
(33) SCHULZE. *Z. Phys.* **78**, 92 (1932).
(34) VON HIPPEL. *Z. Phys.* **80**, 19 (1933); **97**, 455 (1935).
(34*a*). Footnote on p. 477 of second reference (34).
(35) VON GEEL and KERKUM. *Physica*, **5**, 609 (1938).
MAY. *Proc. Phys. Soc.* (in the Press).
(36) TROST. *Phys. Z.* **36**, 801 (1935); *Z. Phys.* **105**, 399 (1937).

§6

(37) CHRISTOPH. *Ann. Phys., Lpz.*, **26**, 145 (1936).
(38) MEDICUS. *Z. Phys.* **103**, 76 (1936).
(39) LEWIS and BURCHAM. *Proc. Camb. Phil. Soc.* **32**, 503 (1936).
(40) DASANNACHARYA and RAO. *Nature, Lond.*, **138**, 289 (1936).
(41) DUNNING and SKINNER. *Rev. Sci. Instrum.* **6**, 243 (1935).
(42) DUFFENDACK, LIFSCHUTZ and SLAWSKY. *Phys. Rev.* **52**, 1231 (1937), and papers cited there.
(43) BOSCH. *Ann. Phys., Lpz.*, **19**, 65 (1934).
(44) EGELHAAF. *Z. Phys.* **108**, 19 (1937).

SOME DIFFICULTIES AND DEVELOPMENTS IN QUANTUM MECHANICS

By H. T. FLINT

§ 1. INTRODUCTION

In a contribution to these *Reports* in 1937 under the title of *The Beginning of the new Quantum Theory* it was shown how great a measure of success had been attained in giving a quantitative account of those phenomena which are described as having their seat within the atom, but outside the nucleus. In this branch of atomic physics the crowning feature is the discovery by Dirac of the first order quantum equation.

In the present contribution we shall consider some difficulties and some other developments of the theory.

The subjects will be *Difficulties associated with the first order Equation*, *The Theory of Radiation* and *Radioactivity*.

The history of physics shows that it is no paradox to say that theories sometimes contribute their greatest service by exhibiting their weak points. When full stops become signs of interrogation research is stimulated and guided and further progress ultimately results.

We are at this stage in the quantum theory at the present time, and this account, though not concerned with the most recent developments, will, it is hoped, show why statements have come to be replaced by questions in some parts of it.

§ 2. THE NEGATIVE ENERGY VALUES OF THE QUANTUM THEORY

The most striking difficulty of the quantum theory can be illustrated by considering the case of a free particle.

In ordinary mechanics a certain amount of kinetic energy is associated with it of magnitude

$$W = \frac{1}{2m} p^2, \qquad \text{.....(1)}$$

where m is its mass and p its momentum.

This quantity can never be negative, in fact the idea of a particle with negative kinetic energy has no place in ordinary mechanics and would appear to be an absurdity.

In relativistic mechanics the expression (1) needs modification. Here the relation between kinetic energy and momentum is

$$E^2 = c^2 (p^2 + m_0^2 c^2), \qquad \text{.....(2)}$$

where E denotes the relativistic expression for the energy.

But although algebraically the value of E may have a positive and a negative value, the theory of relativity rejects the latter and the idea of a particle with negative energy is as absurd in the theory of relativity as in ordinary mechanics. Only the positive value of E plays a part in the theory and the relation between the non-relativistic energy and this positive E is

$$W = E - m_0 c^2. \qquad \text{.....(3)}$$

But the fact that the relation between energy and momentum is quadratic according to equation (2) has a very important consequence in connexion with the solution of the quantum equation.

If reference be made to equation (93) of the article in the *Reports* of 1937, it will be noted that Dirac's equation is a matrix equation. When this equation is written out in the ordinary form it is found that there are four equations between the components of ψ. These components are four in number and since there is no field of force the solutions are simple and of the form

$$\psi_n = A_n e^{2\pi i/h\,(p_x x + p_y y + p_z z - Et)}, \qquad \text{.....(4)}$$

where n has the values 1, 2, 3 and 4, A_n denotes a constant amplitude and p_x, p_y and p_z are components of momentum. The four components of Dirac's equation written out in detail now become[1]

$$\left.\begin{aligned}
(p_x - ip_y)\,A_1 + p_z A_2 + \frac{E + m_0 c^2}{c} A_4 &= 0,\\
(p_x + ip_y)\,A_2 - p_z A_1 + \frac{E + m_0 c^2}{c} A_3 &= 0,\\
(p_x - ip_y)\,A_3 + p_z A_4 + \frac{E - m_0 c^2}{c} A_2 &= 0,\\
(p_x + ip_y)\,A_4 - p_z A_3 + \frac{E - m_0 c^2}{c} A_1 &= 0.
\end{aligned}\right\} \qquad \text{.....(5)}$$

Suppose that a solution has been found for the case when E is positive and that the amplitudes are A_1, A_2, A_3 and A_4. We will write the corresponding values of the components of ψ as ψ_1', ψ_2', ψ_3' and ψ_4', denoting the solution by

$$\psi = (\psi_1', \psi_2', \psi_3', \psi_4'). \qquad \text{.....(6)}$$

If we consider the case of the negative value of E, all we have to do is to write $-E$ instead of E in equation (5), and if we write $\psi_1 = \psi_3'$, $\psi_2 = \psi_4'$, $\psi_3 = -\psi_1'$, $\psi_4 = -\psi_2'$ we return to the set of equations which held for the case of positive E.

Thus if equation (6) is the solution for $+E$, the solution for $-E$ is

$$\psi = (\psi_3', \psi_4', -\psi_1', -\psi_2'). \qquad \text{.....(7)}$$

The rest mass m_0 remains positive and the result indicates that if there is a state with positive energy for a particle of mass m_0, there is also a state with negative energy. It is easy enough to interpret the former state for this is simply the ordinary case of a particle but it is difficult to form a conception of the second state.

We cannot in this case rule out the existence of these negative energy states in an arbitrary manner as in classical and relativistic mechanics.

The complete description of the particle by means of the quantum equation requires account to be taken of all fundamental states.

No particular state can be excluded unless it happens that no transformations take place between it and the other states. If this were the case it would have no physical significance. This is not so for the negative energy state for there exists a definite probability of transition to it from the positive energy state.

The impossibility of excluding it can be seen from a more familiar case. Let us take the example of a stretched string which is capable of an infinite number of states of vibration corresponding to the production of the fundamental and the higher harmonics. Any arbitrary state of the string can be represented by means of Fourier's theorem as a summation of the individual simple states with appropriate amplitudes. These amplitudes are a measure of the contribution made by the particular simple state to the representation. We have to take account of all these simple states and cannot arbitrarily omit any particular one of them. We omit it only when the amplitude associated with it happens to be zero. If we can conceive of a case where a particular harmonic never occurs, we are entitled to rule it out as of no physical importance.

In the quantum theory these amplitudes correspond to probabilities of transition between states and for the same reason no particular one can be arbitrarily disregarded. We are forced to consider the negative energy states seriously for according to the quantum theory they appear to have the right to be considered as physical entities. If the quantum theory is a correct mode of description for the particle with positive kinetic energy, it ought equally to be correct for the particle with negative kinetic energy. We have to consider very carefully how such a particle or rather such a state would appear in the realm of practical physics. It has no meaning in terms of mechanics simply to describe the state as a particle with negative kinetic energy.

Dirac's approach to this question was by means of a very forced hypothesis which cannot claim to be satisfactory but which, nevertheless, led to a very remarkable result[(2)].

It will be remembered that Pauli introduced an exclusion principle which required that a particular state could be occupied by one electron only. Let it be supposed that the negative energy states are almost all occupied. This will mean that transformations to these states are possible only in a few cases.

It has further to be supposed that electrons in these states produce no external field and do not contribute to the energy and momentum of the system. In fact these electrons are equivalent to empty space, the perfect vacuum according to this suggestion being *the region where all states of positive energy are unoccupied and all states of negative energy occupied.*

How will an unoccupied negative energy state appear? Let us see what has to be done to neutralize it. According to the above suggestion we reduce its region to emptiness by placing an electron with negative energy into it. It will itself appear

as a positive electron with positive kinetic energy. It will thus behave like the particle discovered by Anderson in 1932, now known as the positron.

It is a remarkable fact that the suggestion of Dirac was made before the discovery of the positron. At that time no experiment had suggested the existence of a positively charged fundamental particle other than the proton. The idea at first was that the proton might be in some way the counterpart of the empty states but there was no possibility of accounting for the difference in mass.

The suggestion of the existence of a positive counterpart of the electron is no new thing, but often the suggestion had been made on the slender ground of symmetry. Dirac's theory gave to the still undiscovered particle a reality because his suggestion was based on the success of the quantum theory. That success would be seriously in danger if an absolutely impossible state were uncompromisingly proclaimed. If the theory was incorrect on this point it was likely to be incorrect generally. A way out would have been to discover something of the nature of an exclusion principle forbidding the existence of negative energy states. Dirac's suggestion takes the place of this absolute exclusion.

The occurrence of particles with negative energy in a case of reflection of electrons[3]. An application of Dirac's equation to a simple case brings out the difficulty of the negative energy values in a very striking way. The problem which we shall now consider was first examined by O. Klein. We shall simplify the case by supposing that the particles move along the direction of the x-axis. On the left of the plane $x=0$, the potential is supposed to be zero and on the right the potential is electrostatic and is equal to V. Thus we have at $x=0$ a potential discontinuity.

In ordinary particle dynamics the question as to whether electrons will pass across the plane of separation of these regions can be answered by means of an energy consideration.

If they succeed in crossing it, they lose an amount eV of their kinetic energy and if their kinetic energy is insufficient they will not cross at all. But in the quantum theory the question is not answered so simply. As we shall see in the section on radioactivity a certain probability of passage exists even if the kinetic energy is less than eV.

In order to obtain an answer to the question whether the electrons will get through we have to appeal to the quantum equation and find the value of the function ψ.

When an electron exists in a region with no field of force the function ψ has the set of four values given by equation (4) and in the present case with x as the only coordinate we have

$$\psi_n = A_n e^{\frac{2\pi i}{h}(px - Et)}, \qquad \text{.....(8)}$$

where p is the momentum of the particle and E is its total energy. The relation between E and p is given by equation (2).

The electron can now be considered as a wave represented by equation (8) which is incident upon a barrier in analogy with a light wave falling normally upon a sheet of glass. Instead of appealing to dynamics in order to trace the particle we

appeal to the wave theory in order to find the amplitude A_n of the wave represented by ψ_n.

In general we know that the wave will be partly reflected and partly refracted. In the former case the momentum is $p' = -p$ and in the latter we denote the momentum by p''. We shall adopt the notation A, A', A'' or p, p', p'' for the incident, reflected and refracted waves respectively.

In the region to the right of $x=0$, we have kinetic energy $(E-eV)$ and the relativistic value of this is given by

$$(E-eV)^2 = c^2(p''^2 + m_0^2 c^2), \qquad \text{.....(9)}$$

by comparison with equation (2).

The wave equation for this region is

$$\{(E-eV)/c + \alpha p'' + \alpha_0 m_0 c\}\psi = 0 \qquad \text{.....(10)}$$

(cf. equation (93) of last year's *Reports*).

Together with this there is an associated equation written in the form

$$\phi\{(E-eV)/c - \alpha p + \alpha_0 m_0 c\} = 0, \qquad \text{.....(11)}$$

ϕ is a matrix associated with ψ, having likewise four components, and the position of ϕ in front of the equation is required by the convention of matrix operations. α and α_0 are matrices.

We shall not go into the details of the calculations for it will be sufficient to explain the occurrence of negative energy values to quote Klein's results.

We must bear in mind that it is the product of the matrices ψ and ϕ, viz. the expression $(\psi_1\phi_1 + \psi_2\phi_2 + \psi_3\phi_3 + \psi_4\phi_4)$, which gives the probability of the occurrence of a particle at any point.

For the incident wave ψ and ϕ satisfy the equation (10) in which V is zero and the associated equation corresponding to equation (11).

For the reflected wave lying also in the region $V=0$, we replace p by p' and for the refracted wave equations (10) and (11) give ψ and ϕ.

The expression which gives the ratio of the values of $\sum_1^4 \psi_n\phi_n$ for the reflected and for the incident electrons is

$$\left\{\frac{2eVm_0}{e^2V^2/c^2 - (p+p'')^2}\right\}^2.$$

This gives the fraction of the electrons reflected.

When there is no potential barrier, $V=0$, and the value is zero, which means that they all go forward. As V increases the fraction increases and becomes unity when $E - eV = m_0 c^2$.

Beyond this it is evident from equation (9) that p'' is imaginary. But the wave-theory interprets this as total reflection. It corresponds with the imaginary value of the refractive index for a dispersive medium which occurs in ordinary wave theory for a certain range of wave-lengths.

If the ratio of the reflected to the incident electrons is calculated for this region

it is found that the ratio is unity. No electrons pass through the barrier, the solution of the wave equation in this case being a rapidly diminishing exponential function.

This region of total reflection exists as V is increased from the value just given, i.e. $eV = E - m_0c^2$ to the value $eV = E + m_0c^2$. For greater values of V, p' again becomes real and we return to our original expression for the number of particles reflected.

Since the kinetic energy is $(E - eV)$, it is clear that its value is now negative. But although such a value takes us out of the domain of ordinary dynamics, we see that a definite fraction of transmitted particles results from our calculation. The value of the ratio is unity for $eV = E + m_0c^2$, which means that no electrons get through. For values of $eV > E + m_0c^2$ the ratio is less than unity and as V increases its value diminishes to $(E/c - p)/(E/c + p)$, which is the value acquired for very large values of V. Thus the fraction of electrons with negative kinetic energy which pass through the barrier increases to the value $2p/(E/c + p)$ for a very large value of the potential.

This ratio is of the same order of magnitude as the ratio of the velocity of the incident particles to the velocity of light.

The nature of the potential discontinuity in this problem makes it artificial but the high value of the last ratio obtained is not due to the assumption of this particular discontinuity. The theory definitely forecasts these electron jumps even in cases where the potential rises rapidly and continuously.

This example suggests that the theory forecasts a process which has not yet been observed and which appears at present foreign to the ideas of physics.

Bohr first pointed out that it appears that the probability of these transitions acquires finite values when the charge is present in a field in which the potential rises by an amount equal to the rest energy (m_0c^2) in a distance $\hbar/m_0c$. Sauter[4] has shown that this is correct and is in accordance with the application of Dirac's equation to an example of which Klein's problem is a limiting case.

States of negative energy are also of importance in the theory of radiation. Thus in the problem of scattering from electrons in the calculation of the amount of radiation scattered by an electron it is necessary to make use of the idea that the electron is transformed into an intermediate state with negative energy and then from this to the final state with positive energy. In this way the correct scattering formula is obtained.

It may be concluded, and the suggestion has been made, that Dirac's equation is incorrect and that it requires some modification which would remove these difficulties. But Dirac's equation is remarkable on account of its mathematical form. It satisfies the condition required by the special theory of relativity, i.e. it is invariant with respect to the Lorentz transformation and it is in a form which makes it possible to incorporate it without any forced hypothesis in the framework of the general theory of relativity. It is probable that some new limitation to physical theory remains to be discovered or incorporated in the notation which will explain how the difficulties arise[5]. We may, without at this stage insisting too

closely upon the analogy, remind ourselves of the relation of dynamics to electromagnetic theory.

It was in the attempt to combine both these theories that the nature of the velocity of light as a constant imposing a limit upon dynamics was revealed. Ordinary dynamics applies in the limiting case when the velocity of light can be considered as very large. But we can regard this velocity as denoting the limit in the neighbourhood of which we cannot apply our ordinary notions.

It has been suggested that a length of the order of magnitude e^2/m_0c^2, i.e. of the order of the radius of the electron, must be regarded as placing a limit upon our ideas of measurement. It may be that if such a length can be successfully incorporated in our theory the difficulties we have mentioned and others which present themselves may be removed. The research for a way out of the difficulties takes various forms, one way is to try to introduce the electron radius satisfactorily into the electromagnetic theory and at the same time to avoid the difficulty of infinite energy associated with a point charge[6]. Another is to appeal to a geometric conception and to show that the consideration of lengths less than this limit leads to difficulties in space-time description[5]. A third attempt has been based on the assumption that smaller lengths have no physical meaning and some progress has been made in the derivation of a theory from which some of the difficulties we have mentioned disappear.

§ 3. THE THEORY OF RADIATION[7]

The quantum theory began with Planck's theory of radiation and it was not until some years later that a quantum theory of mechanics was developed. In modern quantum theory the latter aspect has been developed to a high degree leaving the theory of radiation somewhat behind. The procedure is somewhat similar to that of Planck's theory. It will be remembered that he made use of the idea of the harmonic oscillator in the development of his theory of radiation. To-day the quantum theory of radiation is based on the quantum theory of a particle and the first step is to reduce the classical theory of the radiation field to the form of generalized dynamics for it is on this form that the quantum theory is based.

By the quantum theory of radiation we mean a theory based upon quantum principles which will account for interference phenomena and for the processes of emission and absorption. Dirac was the first to succeed in constructing such a theory and we shall see that it is based upon a very simple physical concept. When we come to the consideration of the case where electric charges interact we find that difficulties arise, some inherited from the classical conception of the electron but others peculiar to the quantum theory itself.

The field of pure radiation. First we consider the case of a region containing radiation alone with no particles of any kind. This field of radiation can be described in the familiar way by Maxwell's electromagnetic equations and it offers the simplest example of Maxwell's theory. Quantities such as field intensities and energy per

unit volume are all derivable from four quantities, viz. A_x, A_y, A_z and ϕ. These are the three components of the vector potential **A** and the electrostatic potential. When there are no charges present these four components can be reduced formally to three and we can refer the various quantities to three components of vector potential, omitting ϕ. We shall still represent these components by A_x, A_y, A_z, and the vector by **A**. The two equations which this vector satisfies are

$$\nabla^2\mathbf{A} - \frac{1}{c^2}\frac{\partial^2\mathbf{A}}{\partial t^2} = 0, \qquad \text{.....(12)}$$

$$\operatorname{div}\mathbf{A} = 0. \qquad \text{.....(13)}$$

If we consider a closed region of space and assume certain boundary conditions, it is possible to expand **A** in the form

$$\mathbf{A} = \sum_s \mathbf{e}_s u_s \sin\left\{\frac{2\pi\nu_s}{c}(l_s x + m_s y + n_s z) + \alpha_s\right\}. \qquad \text{.....(14)}$$

u_s is a function of the time only and the representation means that the electromagnetic field is represented by the superposition of plane waves. $\mathbf{e}_s$ is a unit vector and (l_s, m_s, n_s) are the direction cosines of a component plane wave. Since in the pure radiation field we are concerned with transverse waves the vector $\mathbf{e}_s$ and the direction (l_s, m_s, n_s) are at right angles.

The purpose of this report being to give the broad outlines of the principles and methods of the procedure, we shall omit the details of calculations.

The expression for the energy in the field is

$$U = \frac{1}{8\pi}\int (E^2 + H^2)\, dv, \qquad \text{.....(15)}$$

and it can be shown that the average per c.c. is

$$W = \frac{1}{8\pi c^2}\sum (\tfrac{1}{2}u_s^2 + 2\pi^2\nu_s^2 u_s^2). \qquad \text{.....(16)}$$

If we write $u_s = \sqrt{(8\pi c^2)}\, p_s$, $u_s = \sqrt{(8\pi c^2)}\, q_s$,

$$W = \sum (\tfrac{1}{2}p_s^2 + 2\pi^2\nu_s^2 q_s^2). \qquad \text{.....(17)}$$

This is at once recognized as being in the form of the sum of the energies of simple harmonic oscillators. Moreover if equation (14) be substituted in equation (12) we find

$$\ddot{q}_s + 4\pi^2\nu_s^2 q_s = 0, \qquad \text{.....(18)}$$

which is the equation of motion of a simple harmonic oscillator with a frequency ν_s.

In this way the field becomes the seat of oscillators of frequencies ν_1, ν_2, ν_3, etc., and since we know how to quantize the motion of these oscillators, we can, through them, quantize the field.

We can, for example, obtain an expression for the energy of the field in a particular quantized state.

The procedure is to replace the familiar mechanical quantities p_s and q_s by the

corresponding matrices of the quantum theory. We use the same symbols for these matrices remembering that they satisfy

$$p_s q_s - q_s p_s = h/2\pi i, \qquad \ldots\ldots(19)$$

$$p_s q_r - q_r p_s = 0 \quad (r \neq s). \qquad \ldots\ldots(20)$$

The values which these assume can be illustrated by those for q_s.

The matrix components are of the type

$$q_{n,n+1} = \sqrt{(\tfrac{1}{2}h\,(n+1))}\; e^{-2\pi i r t}. \qquad \ldots\ldots(21)$$

Such components as $q_{1,3}$, $q_{2,4}$, etc. vanish, only those of the type $q_{n,n+1}$ and $q_{n,n-1}$ exist.

The energy of the quantized particles and thus of the field of radiation is

$$W = \sum_s n_s h\nu_s. \qquad \ldots\ldots(22)$$

The quantum state of the field is described by the quantum states of the oscillators, i.e. by the numbers n_s. Transformations in the field are described by a change from one set $(n_1, n_2, n_3, \ldots)$ of quantum numbers to another set $(n_1', n_2', n_3', \ldots)$.

Another important quantity is the momentum of the field

$$\mathbf{G} = \frac{1}{4\pi c}\int \mathbf{E} \times \mathbf{H}\, dv, \qquad \ldots\ldots(23)$$

where $\mathbf{E} \times \mathbf{H}$ denotes the vector product of $\mathbf{E}$ and $\mathbf{H}$.

On substituting for $\mathbf{E}$ and $\mathbf{H}$, which are derived from equation (14), we obtain

$$\mathbf{G} = \Sigma \mathbf{G}_s = \sum_s n_s \frac{h}{c} \nu_s \mathbf{r}_s, \qquad \ldots\ldots(24)$$

where $\mathbf{r}$ is written for the unit vector in the direction (l_s, m_s, n_s).

Both energy and momentum depend upon $h\nu_s$ and it is as if we had, corresponding to the state n_s, n_s quanta of energy and the same number of quanta of momentum.

This presentation of the theory of radiation is strongly reminiscent of Planck's original assumption concerning the oscillators he introduced to assist in the derivation of his formula. We have here a consistent notation for the expression of that assumption.

The field containing radiation and electric charges. In this case the total energy will consist of the energy of the oscillators which represent the radiation (cf. equation (17)), the energy of the charges and a term representing a coupling between the oscillators and charges. The last of these must appear since we know that an interaction between electric charges and radiation occurs. In order to be able to appreciate the procedure in this case let us return to equation (2). This is an expression for the energy when there is no electromagnetic field. In this case when we pass from ordinary dynamics to quantum dynamics p, the ordinary momentum is replaced by an operator. This procedure rests on the fact that the components of momentum (p_x, p_y, p_z) are conjugate in the Hamiltonian sense to the coordinates

(x, y, z). In the presence of a field these components are no longer conjugate to (x, y, z) but are replaced by quantities (Π_x, Π_y, Π_z) related to them by

$$\Pi_x = \left(p_x + \frac{e}{c} A_x\right), \text{ etc.}$$

The total energy E for a single charged particle is now $(mc^2 + e\phi)$ and the energy W occurring in non-relativistic theory is related to E by equation (3). Note that m is the ordinary mass and m_0 the rest mass.

Equation (2) is equivalent to

$$m^2c^2 = p^2 + m_0^2c^2,$$

and now becomes

$$\left(\frac{W - e\phi + m_0c^2}{c}\right)^2 = \sum_{x,y,z} \left(\Pi_x - \frac{e}{c} A_x\right)^2 + m_0^2c^2. \qquad \text{.....(25)}$$

In passing into the quantum theory Π_x, Π_y and Π_z must be replaced by operators. It will be sufficient for the purpose of illustration if we neglect the terms in $1/c^2$, when we obtain

$$W = \frac{1}{2m_0} \mathbf{\Pi}^2 + e\phi - \frac{e}{m_0c} (\mathbf{A}.\mathbf{\Pi}), \qquad \text{.....(26)}$$

where $\mathbf{\Pi}$ is written for the vector with components Π_x, Π_y, Π_z and $\mathbf{A}.\mathbf{\Pi}$ denotes a scalar product.

Remembering the expansion of $\mathbf{A}$ in equation (14) and adding equations (17) and (26) we obtain for the energy of the oscillators together with that of the charge

$$H = \frac{1}{2m_0} \mathbf{\Pi}^2 + e\phi - \frac{e}{m_0c} \sqrt{(8\pi c^2)} \sum_s (\mathbf{A}_s.\mathbf{\Pi})\, q_s \sin\theta_s + \sum_s (\tfrac{1}{2} p_s^2 + 2\pi^2 \nu_s^2 q_s^2), \qquad \text{.....(27)}$$

where θ_s is written for the argument of the sine function in equation (14).

The appearance of the third term on the right in the expression for the energy is very important. It contains terms referring to the radiation and to the charge and represents the coupling that exists between them. Without it we are left with contributions from both but there is no possibility of interaction, the two systems remaining isolated.

As we have stated above, Dirac based his theory of radiation on a simple idea. If we have two oscillating systems with no connexion they carry on their motion independently. But if the two systems are coupled by a very light system a slight perturbation occurs. If one system is initially at rest it will be affected by the other to a greater or less extent according as the frequencies of the two systems lie close to one another or not. The total energy will contain a term expressing this coupling.

In the problem before us the radiation and the charge are considered as vibrating systems lightly coupled. For simplicity we have considered the case of one charge only. If more are present we have to add a corresponding term for each.

The simple harmonic equation of motion (18) is now modified and it can be

shown by the methods of classical dynamics that the corresponding equation in this case is

$$\ddot{q}_s + 4\pi^2\nu_s^2 q_s = \frac{e}{m_0}\sqrt{(8\pi)}\sum_r (\mathbf{A}_r.\mathbf{\Pi})\sin\theta_r, \qquad \text{.....(28)}$$

for the sth oscillator.

From these equations, which are now in the Hamilton form, we can obtain the energy operator corresponding to H in equation (27). It is upon this operator that the solution of the problem depends.

If we proceed according to Schrödinger's method we obtain a number of results in agreement with experiment but we are then dealing only with problems in which the relativity correction is neglected.

In order to discuss the problems of radiation completely we must use the method which depends upon the use of the first order quantum equation. The method of solution in either case is the method of perturbations. The systems are supposed to be known separately and from the functions describing them we obtain solutions applicable to the case when they interact through a light coupling.

In this way we obtain a satisfactory theory of emission, of light propagation in empty space and of Lippmann's interference to mention only three as examples.

We shall mention also the case of scattering of radiation by free electrons because we wish to refer to it later on.

In order to discuss this problem it is necessary to go to a higher order of approximation than that represented by equation (26) for the perturbation term in that approximation makes no contribution to the scattering.

Thus we have to look for the coupling in the next order of approximation. This provides an energy term H' and in the quantum theory this is a matrix with components which represent the energies of transitions between different states of the charge and radiation.

By means of this matrix we deduce certain quantities a which may be described as amplitudes and the squares of which give probabilities of transitions between states. If the radiation field is in a quantum state represented by the states of the oscillators whose quantum numbers are n_1, n_2, etc., and with a single charge in the quantum state n, then the probability of that state is represented by

$$(a_{n\,n_1\,n_2\,\ldots})^2.$$

In order to illustrate the use of this notation in the case of scattering let it be supposed that initially the radiation field is represented by an oscillator numbered s in the state n_s with all other oscillators in the zero state and with a single charge in state n. Since this is certainly the initial state we begin with

$$a_{n\,0\,0\,\ldots\,n_s\,0\,\ldots} = 1. \qquad \text{.....(29)}$$

We require to find the intensity of radiation which after interaction with the charge lies in some other state. Let us take the simple case when the charge remains in the same state but that one quantum of radiation leaves the state n_s and enters a state labelled r.

The energy in n_s is $n_s h\nu_s$, and if one quantum leaves it the energy becomes $(n_s-1)\,h\nu_s$ and the state becomes (n_s-1).

The state r now possesses energy $h\nu_r$ and the state is 1_r.

Thus we require the amplitude $a_{n\,0\,0\,\ldots\,n_s-1\,0\,0\,\ldots\,1_r\,0\,\ldots}$.

The theory which we have outlined actually gives this value.

If we are concerned with the probability of the scattering of a single quantum we have to remember that this may leave the scattered quantum in any state r and the total probability of scattering is obtained by integrating over all possible final states.

If we limit ourselves to a short time the value of the probability depends on the integral

$$\int_0^\infty \frac{\sin^2 \pi (\nu_s-\nu_r)\,t}{(\nu_s-\nu_r)^2}\,d\nu_r.$$

The contribution to this integral comes from the region close to $\nu_r=\nu_s$, and this means that there is only a very small probability of scattering unless $\nu_r=\nu_s$. In other words the scattered frequency is the same as that of the incident radiation.

Expressed in terms of energy we can say that the probability of scattering is appreciable only when the interaction takes place in accordance with the principle of conservation of energy for in the case considered $h\nu_s=h\nu_r$ and no change occurs in the state of the charge. We have considered this case chiefly to bring out this feature of the problem which shows how conservation of energy appears in certain problems of the quantum theory. It should be noted that we are not compelled to provide for it in any special way, it is included naturally within the notation.

This very simple example of scattering leads to the formula of J. J. Thomson, viz. the scattered energy in a short time t is

$$\frac{8\pi}{3}\frac{e^4}{c^2m^2}\,n_s h\nu_s t \text{ per c.c.}$$

When Dirac's equation is used in the calculation of the perturbation effect we obtain the same final result if we approximate to the same degree. There is a fundamental difference in that in this method the perturbation term is linear and not quadratic in the electromagnetic potential.

It turns out that no transitions occur between states in which more than one oscillator changes its quantum number, e.g. in the above case where n_s becomes n_s-1 and 0_r becomes 1_r.

This seems to mean that transitions do not occur at all in the case we have considered and that in consequence there would be no scattering. That is definitely incorrect but the theory has something more to say on the point, for although these direct transitions do not occur we can arrive at the same final result if they occur in stages through some intermediate state.

According to the theory these intermediate states are those for which the electron has negative energy. The scattering formula arrived at by this process is again the Thomson formula and here once more we have an example of the importance of the negative energy states and of the necessity of retaining them in a correct theory.

This case of scattering serves to illustrate the general method used in the problem. Another approximation gives an account of Compton scattering and the exact theory gives the result obtained by Klein and Nishina.

Quantum electrodynamics. The theory which we have outlined is not a general theory of the electromagnetic field for we have considered it by the superposition of plane waves. As we know from the electromagnetic theory div E vanishes for this case, where E denotes the electric intensity. But in the general field theory we have

$$\text{div } E = 4\pi\rho, \qquad \ldots\ldots(30)$$

where ρ is the electric density.

In spite of this, it is possible to reduce the field equations to a mechanical form and from the mechanical Hamiltonian to obtain the corresponding quantum operator. In this process equation (8) is replaced by three equations of the same type with three sets of variables instead of the single set of which q_s is typical. These equations represent the fact that in the general electromagnetic field the vector potential is considered as compounded of three components, one in the direction of the wave and two transverse.

In addition to these there is another equation of the same form containing a coordinate which is introduced in the expansion of the electrostatic potential corresponding to u_s of equation (14).

In this way a consistent scheme for the treatment of the problems of radiation is set up and the theory we have considered is included within it.

But a difficulty arises at once. It is, however, a difficulty belonging to the classical theory and results from the fact that so much of that theory is taken over in founding a system of quantum electrodynamics. A group of terms in the Hamiltonian function represents the electrostatic energy due to the mutual action of the changes. This energy is given by the familiar expression $\frac{1}{2}\Sigma \frac{e_m e_n}{r_{mn}}$, where e_m and e_n denote the charges and r_{mn} their distance apart.

The difficulty is that the summation must include the term e_m^2/r_{mm}, where r_{mm} approaches zero. This is merely the statement that the self-energy of a point charge is infinitely great.

In the classical theory the difficulty has been avoided by giving a small but finite radius to the fundamental charge—the electron.

Although the introduction of a length of the order e^2/m_0c^2 is foreign to the classical theory no difficulty of a fundamental character is raised. In the theory we are discussing the fundamental charge is a point charge and from this point of view the conception of an electron with a structure and an associated radius has no significance. But the infinite value of the energy term persists and the theory can do no better than ignore it in spite of the fact that invariance with respect to the Lorentz transformations is destroyed after much trouble has been taken at an earlier stage to preserve it.

But from what has been said on this point we cannot lay any special blame upon the quantum theory for this difficulty.

The classical theory concealed this skeleton in the cupboard, the quantum theory has not hesitated to bring it out into the full glare of daylight.

But if in this particular case the younger theory is allowed to share the guilt with its predecessor, it must bear full responsibility for another and equally serious difficulty.

In the theory of radiation the interaction term of equation (26), translated into its quantum form, is treated as a small perturbation and higher degrees of accuracy are determined by successive approximations. In the classical theory the higher terms depend upon the structure of the electron and are vanishingly small for a point charge.

In the quantum theory, however, the higher approximations diverge even in the case of a point charge. It is therefore presumed that this difficulty is inherent in the process of quantization.

The conclusion is drawn from the success of the theory on the one hand and the difficulties that arise on the other that it is satisfactory in those problems in which the structure of the electron is not concerned, but in any case where internal properties of this fundamental charge arise no satisfactory solution is available.

Some new physical ideas are needed.

§ 4. THEORIES OF RADIOACTIVITY[8]

In this section we shall discuss the theories of alpha- and beta-ray disintegration. The attempt to find a satisfactory theory in the case of the latter is still going on and the difficulties which are disclosed here have given both stimulus and direction to the lines of modern research in nuclear physics. We shall concern ourselves with a very remarkable theory put forward by Fermi.

This is by no means completely satisfactory and other methods of approaching the problem are available, but in this report where it is impossible to enter very extensively into the whole problem we limit ourselves to Fermi's theory because it is of fundamental importance both in the history and in the development of the subject.

The theory of alpha-ray disintegration. This theory was first given by Condon and Gurney and by Gamow who published their work independently and almost simultaneously.

This treatment of the problem may be described as one of the most elegant and satisfactory contributions to physics within modern quantum theory. The statistical character of the theory finds its natural expression in this problem and the notation is particularly adapted to the observations it portrays. Moreover it is in this case that we obtain an early example of our need to modify our mechanical ideas in the consideration of nuclear problems. The subject to be treated is essentially the passage of particles through potential barriers. The nucleus is for the purpose of the calculations a region within a potential boundary. In this region we suppose that alpha particles exist and from which they can escape, not according to the laws of ordinary mechanics but according to those of quantum mechanics.

The problem is similar to that of Klein's example discussed earlier in this report. But here the case is simpler for we can omit the relativity correction and can therefore make use of Schrödinger's equation and need not apply Dirac's equation. The difficulties in Klein's problem do not arise.

In order to appreciate the difficulty which faced any explanation of the phenomenon of radioactive decay on the basis of a rigid mechanical theory we may take the example of emission from uranium.

Scattering experiments showed that the Coulomb law of force held to within a distance of $3{\cdot}2 \times 10^{-12}$ cm. of the nucleus so that it was possible to estimate the shape of the potential barrier to this distance. The energy of the alpha particle measured experimentally was less than that required to cross this barrier and in fact was equal to that possessed at $6{\cdot}3 \times 10^{-12}$ cm. from the nucleus.

Thus according to the ordinary conception of particle dynamics it was impossible for the particle to escape.

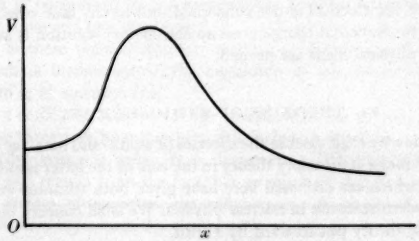


Figure 1.

Klein's problem has shown us that there is no difficulty from the point of view of quantum mechanics. According to this, if there is a probability that the particle lies within the barrier there is a probability that it will escape which may be different from zero.

A curious feature of radioactive elements is the great differences shown in their rates of decay, ranging from a fraction of a second to millions of years without any apparent extraordinary difference between the nuclei.

Condon and Gurney showed that the value of the decay constant, the reciprocal of which gives the average life of the element, depended upon a certain integral and that on this account the variation in the periods could be accounted for.

Finally the law of Geiger and Nuttall, which relates the decay constant and the energy of the particle emitted, could be developed from the theory.

Let us suppose that the potential V is a function of x, the distance from some convenient origin.

We may suppose that the graphical representation of the function is that of figure 1.

The problem is to discover what chance a particle lying to the left of the potential maximum has of passing to the right of it.

The answer to this question in terms of ordinary mechanics is simple. If the particle has sufficient energy to ride over the barrier it will do so, if not it will always remain on the left.

But expressed in terms of the quantum theory we ask with what amplitude will a wave lying to the left of the boundary escape to the right of it. The wave in question is the representation of the particle and it expresses an aspect of the particle not contained in the ordinary mechanical theory. But in this problem we do not depart very far from the particle conception. We think of a body first on one side of the barrier and then on the other. Its individuality is preserved and the difficulty from the particle point of view is its nature within the barrier within which it ought not to penetrate according to the particle representation. This point is mentioned because of a contrast with the case of beta-ray disintegration where we cannot keep so closely to the idea of a particle which preserves its identity throughout.

The problem can be solved when the shape of the potential barrier is known, i.e. when V is known as a function of x.

The actual case before us is not a simple linear problem but we assume spherical symmetry about the nucleus. After a short distance V becomes $\propto 1/r$, representing the Coulomb type of field. But within some distance $r=r_0$ there is a deviation from this form and in order to obtain a form of potential which enables the Schrödinger equation to be solved and at the same time does not depart too far from what may be supposed to be the actual state of affairs, we suppose that within the distance r_0 the potential is constant, $V=V_0$.

If the number of the nucleus is Z, the Coulomb potential of the alpha-particle is $2(Z-2)e^2/r$.

As we have already mentioned a suggestion for the form of the solution comes from optical theory. In this case the refractive index formally assumes a complex value and this is interpreted as an exponential falling off of the wave. In the present case it is the energy which unites with an imaginary term to form what may be described as the complex energy value.

If this be written in the form $\left(E+\frac{h}{2\pi i}\frac{\lambda}{2}\right)$, the function ψ contains the factor $e^{-2\pi i(Et/h-\lambda t/2}$, E denoting the energy value and λ the rate of falling off of the square of the amplitude of the wave function. It is in fact the disintegration constant.

From the characteristic solutions of the wave equation both E and λ are obtained. The latter contains the exponential factor $\exp\left[-\frac{4\pi}{h}\int\sqrt{\{2m(E-V)\}}\,dr\right]$ where the integration is made across the potential barrier. The relation between the disintegration constant, energy and potential barrier is thus precisely stated.

Condon and Gurney determined the values of λ by plotting the integrand against r and determining the integral graphically.

It was in this way that they were able to account for the wide range in the value of λ.

Further the relation between this quantity and the energy of the emitted particle which the theory gives was found to be the quantitatively derived expression of the well-known Geiger-Nuttall law.

The theory of beta-ray disintegration. The phenomenon of beta-ray disintegration is more complicated than that of alpha-ray disintegration. In the case of the latter the particles emitted are to a high degree homogeneous, although there are exceptions to this or in some cases definite groups occur.

In the case of the former there is a discontinuous distribution of energy in the particles emitted by an element.

Thus figure 2 represents the general form of the relation between the number of particles emitted and the energy possessed by them in the case of Ra E, although it is not to be supposed that the curve is so well known as the figure appears to indicate. The beginning and end cannot be drawn with certainty from the results of experiments.

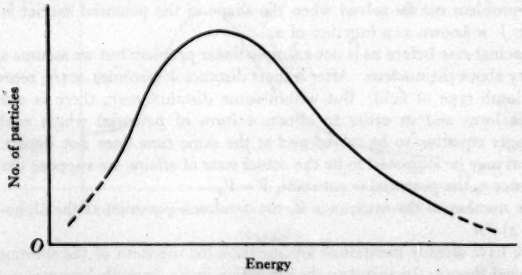


Fig. 2

It might be thought that, as in the case of the alpha particles, the beta particles are emitted by a nucleus which begins and ends in energy states with definite energy. The spectrum could then be explained on the assumption that the particles lost energy in getting away from the atom.

This is a question which has been answered by an appeal to a calorimetric experiment carried out by Ellis[9]. In association with Wooster he measured calorimetrically the energy emitted by Ra E. This element emits very little energy in the form of gamma rays and is thus very suitable for the study of the total beta-ray emission. If the energy emitted is divided by the total number of particles emitted, the average energy per particle is obtained. The actual value obtained agrees with the average energy calculated from the spectral distribution.

If the particles are emitted all with the same energy we expect to find the average value equal to or very nearly equal to that of the maximum energy that the particles are observed to possess.

There is no doubt as a result of this experiment that the latter is not the case. We thus suppose that different atoms of an element emit these particles with different energies. The natural conclusion from this would be that different nuclei of the same element remain with different energies after beta-ray emission. But if this were so, other nuclear processes such as emission of alpha particles would be different and this is not observed. We have to accept the fact that the nuclei, whether they emit beta particles with high or low energies, are all identical. It is to be observed that what is really troubling us here is the apparent break-down of the principle of conservation of energy. Identical nuclei all in the same state lose different amounts of energy and become again identical nuclei apparently all in the same state. In fact, if we wish to consider subsequent nuclear reactions, we have to ascribe to them the same value of the energy.

Bohr[10] suggested that possibly the principle of conservation of energy does not hold in these sub-atomic processes. Certainly we cannot deny this on the basis of direct experiment and it may be that we are making a mistake in holding fast to the mechanical picture of the process. But the principle of conservation of energy is so far reaching in physics and cannot lightly be put aside, moreover there are other difficulties of a fundamental character associated with abandoning it in this case[11].

For these reasons a suggestion of Pauli's has been more acceptable and it has the advantage that it has been fruitful in forming the basis of a theory of beta-ray disintegration. The suggestion is that in the process a neutral particle called the neutrino is ejected with the electron. The two together take away the same amount of energy from a nucleus in each case. Unfortunately it does not appear possible to observe the neutrino, in any case it has certainly not been observed yet, but the problem of finding the energy of the beta particle becomes that of finding how a given amount of energy is divided between these two particles.

There is another difficulty of a mechanical character that concerns the principle of conservation of angular momentum which makes the presence of a second particle equally necessary. This difficulty arises from the fact that it is necessary to ascribe to nuclei of even mass numbers spin momentum of an amount equal to an even multiple of $h/4\pi$ whereas in the case of nuclei of odd number this amount is an odd multiple of this quantity. In beta disintegration there is no change in mass number but the emitted particle carries away a spin momentum of amount $h/4\pi$. This raises a difficulty just as striking as that of the energy difficulty and the presence of a second particle removes it, if it also has a spin of $h/4\pi$.

We shall end this report by giving a short account of Fermi's theory which succeeds to some extent in developing a formalism fitted to this problem and in accounting for its general features.

This theory is based on an analogy with radiation. The emitted particles are like the quanta emitted from an atom when electrons make their jumps from one level to another. This problem of the emission of quanta is not based on the idea that quanta initially present in the system are subsequently emitted by it. They come into being as a result of a change of state in the outer layers of the atom. In the same way we do not now think of the electrons and neutrinos as initially

present in the nucleus, they are generated in the act of disintegration by a change of state in the nucleus. We have thus to make use of some idea on which to base this nuclear change of state and we have to maintain the principles of conservation of electric charge, of energy and of momentum.

It is quite in accord with quantum principles to suppose that electrons do not exist as individual particles in the nucleus. It is in fact impossible to maintain the particle conception of an electron situated within the space of the dimensions of the nucleus. The nucleus is assumed to consist of heavy particles, i.e. of protons and neutrons. This assumption agrees with the ideas of the quantum theory and is proving to be both satisfactory and fruitful in nuclear theories. Moreover it is assumed to be possible for a proton to change its state and become a neutron and vice-versa. When the transition occurs the charge is conserved by the appearance or disappearance of an electron. We shall not speak here of the appearance of a positron and antineutrino which in Fermi's theory can be regarded as formally the same as the disappearance of electron and neutrino.

The notation has to represent the transition from the proton to the neutron state and the associated appearance of an electron in state s (say) and a neutrino in state σ (say) and alternatively the transition from the neutron to the proton state with absorption of an electron and neutrino in their respective states.

The method adopted by Fermi is to construct a Hamiltonian for this process and to regard it as containing a perturbation term like the Hamiltonian representing the interaction between oscillators and charges in the radiation theory.

It will be remembered that in the radiation theory we wrote down the Hamiltonian for the oscillators alone, for the charges alone and added the perturbing Hamiltonian.

The oscillators represent the radiation so that in Fermi's theory the oscillators become the light particles, i.e. electrons and neutrinos. The charges in the radiation field correspond to the heavy particles, i.e. protons and neutrons.

Thus to carry out the analogy we have now to write down the Hamiltonian for the heavy particles alone, that for the light particles alone and add a perturbing Hamiltonian to represent the generation of electrons and neutrinos according to the procedure we have described.

For the Hamiltonian of electrons and neutrinos the solutions giving the ψ-functions are taken from Dirac's equation since it is clearly essential in the case of these fast particles to take account of the relativistic effect.

The relativistic form of the expression for the heavy particles is not known but Fermi succeeded in constructing an expression for the interaction by analogy with the radiation formula and obtained the probability of transitions by the method of perturbation which was applied in the case of radiation. As we should expect by this analogy, the principle of conservation is automatically satisfied by this procedure and it is found that the probability of the transition is only appreciable when the energy of the light particles together with that of the heavy particles remains unchanged.

In this form of the theory the Hamiltonian is made up of the functions ψ and

not of their derivatives. A later modification by Konopinski and Uhlenbech made use of a modified form containing derivatives of these functions.

The result of this work has been to give a formula for the probability of the transition resulting in the creation of a beta particle with a certain energy which resembles the form of the beta-ray spectrum.

The upper limit of the curve corresponding to the high energy values is very interesting. The course of the curve at this limit appears to indicate that the rest mass of the neutrino is very small.

The difficulties of putting the theory to a quantitative test are great on account of difficulties of determining experimentally the accurate form of the spectrum. We shall not go into the question of the quantitative accuracy of the theory since we are concerned to indicate the general traits of this theory of radioactive decay and to show how they portray its features by what is in principle a very simple physical analogy. That is the chief contribution of Fermi's remarkable theory.

REFERENCES

(1) Frenkel. *Wave Mechanics, Advanced General Theory*, p. 278 (Oxford, Clarendon Press).
(2) Dirac. *Quantum Mechanics*, 2nd ed. p. 270 (Oxford, Clarendon Press).
(3) Klein, O. *Z. Phys.* **53**, 157 (1928).
(4) Sauter. *Z. Phys.* **69**, 742 (1931).
(5) Heisenberg. *Ann. Phys., Lpz.*, **32**, 20 (1938).
March. *Z. Phys.* **105**, 620 (1937).
Flint. *Proc. Roy. Soc.* A, **159**, 45 (1937).
(6) Born. *Proc. Roy. Soc.* A, **143**, 410 (1934).
Born and Infeld. *Proc. Roy. Soc.* A, **144**, 425 (1934).
(7) Fermi. *Rev. Mod. Phys.* **4**, 87 (1932).
Dirac. Chapters on Radiation, see reference (2).
Heitler. *Quantum Theory of Radiation* (Oxford, Clarendon Press).
(8) Condon and Gurney. *Nature, Lond.*, **122**, 439 (1928); *Phys. Rev.* **33**, 127 (1929).
Gamow. *Z. Phys.* **51**, 204 (1928).
Gamow and Houtermans. *Z. Phys.* **52**, 476 (1928).
Gamow. *Atomic Nuclei and Nuclear Transformations*, chap. v (Oxford, Clarendon Press).
(9) Ellis and Wooster. *Proc. Roy. Soc.* A, **117**, 109 (1927).
(10) Bohr. *Convegno di Fisica Nucleare* (Rome, 1933).
(11) Landau. Solvay Congress, Discussion on β-disintegration (1934).
(12) Fermi. *Z. Phys.* **88**, 161 (1934).
Gamow. *Atomic Nuclei and Nuclear Transformations*, chap. vii (Oxford, Clarendon Press).
v. Weizsäcker. *Die Atomkerne*, chap. iv (Akad. Verlags, Leipzig).

THE TEACHING OF PHYSICS IN SCHOOLS

By A. W. BARTON

§ 1. INTRODUCTORY

In considering the teaching of physics in schools it is important to realize the difference in atmosphere between a school and a university, and as I had the good fortune to teach for some three years in a university before going to a school, I may claim to know a little of that difference. In the first place, a school is an average cross-section of the community; its population contains a few clever boys, a considerable number of boys of average intelligence and a few boys with little or no aptitude for book learning, but who may have gifts which can be of service to the community, such as personality, which will make the boy into a good Army officer, or judgment of the tastes of his fellow-men, which will ensure his success in business. A university, on the other hand, consists of the intellectual cream of the schools, the remainder having either no inclination or not enough ability to qualify for entry. It consists of the intellectual aristocracy of the community. A university is organized into separate departments, which have their own buildings and have not much contact with the work of other departments. In the physics department, the importance of physics is taken for granted; the students have a reasonably keen interest in it or expect to get their living out of it; there is a keen interest in the results of the latest work, the value of research is axiomatic, and the importance of physics to the community is an accepted fact. A very different state of affairs obtains in a school, which works as an organic whole with close contact between the different branches of knowledge. In an average sized school containing 400 boys there may be at any one time twenty boys who are keen on physics, of whom one or two may go on to the university and take a special honours degree with outstanding success. There will be an equal number who are quite indifferent to the subject or are positively repelled by it. They can make nothing of it; they cannot see its significance in everyday life and they hate its insistence on cause and effect and on the supremacy of law in Nature. The rest of the boys in the school will be of average intelligence, that is, well below the average of the university student, and have no special aptitude for or interest in physics. This, then, is the material to which physics is taught in schools, a cross-section of the population, not the intellectual cream. Then there is a second important difference between the population of a school and that of a university. The undergraduate is reasonably mature; he has some sense of the abstract, he can appreciate the idea of a line of force as the direction of the magnetic field quite apart from the presence of any concrete iron filings to materialize the direction; his logical faculty is beginning to develop and so he naturally thinks of an electric current as a flow of the substance on an ebonite rod which attracts pieces of paper and welcomes some explanation of the phenomena along these lines; his difficulties are to some con-

siderable extent the difficulties of his teachers. How different is the mental state of the boy of fourteen! He has a vivid sense of the concrete, but little if any grasp of the abstract; if he is shown a simple cell lighting an electric lamp, he is fascinated by the mysterious, even romantic, way in which the light is produced; he sees the light, he sees the bubbles of gas coming off the copper plate, he perceives the zinc dissolving. But ask him why all this is happening and he is not very interested. Talk to him about the chemical energy released by the dissolution of the zinc, its conversion into electrical energy and its conversion into heat and light in the lamp, and he is quite unmoved by the logical beauty of the sequence of energy transformations going on before his eyes. His logical faculty has not developed sufficiently to appreciate it; he does not realize that there is anything to be explained. It is the facts that interest him primarily. Again he cannot yet distinguish essential facts from matters of detail. One way of trying to explain the difference between temperature and quantity of heat is to discuss what happens when a piece of hot iron, for example, is plunged into a bucket of cold water. If a boy is asked what happens, he will reply that the water sizzles and only when he is asked what happens to the temperature of the water and iron will he add that the water warms up and the iron cools down. And he will answer the question in a bewildered way indicating that the point is of little interest to him. It is the visually striking facts which impress themselves on his mind. He has no logical faculty to enable him to select the essential facts and discard trivialities. Again the boy's mind is crude and has not yet had the long training needed before it can appreciate the small but pregnant differences of such importance in scientific progress. To show him that a piece of iron becomes a magnet when it is placed near to another magnet, it is not a good thing to do it with an ordinary sized bar magnet, as the iron picks up so few iron filings, that the boy is never quite convinced in his own mind that they are not just sticking on even though a control experiment is carried out. But he will be convinced if a good electro-magnet and a fair-sized piece of iron are used so that it will pick up a box full of nails when the magnet is excited. Lastly such logical faculty as the boy has expresses itself at this stage by an interest in machinery; like Maxwell he wants to know "what is the go of it", but that is all. To sum up, then, there are two differences between a school and a university; the university has the intellectual cream of the country, while the school has much more nearly the average citizen; the university undergraduate has travelled a good way along the road to maturity and has a well-developed logical sense, while the schoolboy is still immature and has a vivid sense of fact and of the concrete, an excellent memory, but a rudimentary sense of logic.

It is the purpose of this article to discuss why physics is taught in schools. What are the aims of such teaching? What methods are being used to realize those aims? To what extent have they been realized and what modifications are needed in order that they may be attained? This will be done by giving a short historical survey, followed by an account of the present position both as regards elementary and advanced work in schools, concluding with some suggestions for future modifications in the present position.

§ 2. THE HISTORY OF THE TEACHING OF PHYSICS IN SCHOOLS

The only schools which existed in the middle of the nineteenth century were the public and grammar schools. They were isolated from one another, but the feature they had in common was their contact with Oxford and Cambridge and with the newer universities which were springing up. It is fair to say that the majority of boys who went to those schools did so in the hope of going to the university afterwards; they were, in fact, gateways to the universities. It was therefore natural that they should provide courses of study suitable for pupils who were afterwards to go to the university; their courses were designed for future specialists or professionals. At that time most of the scientific work in this country was being done by aristocratic amateurs, although there were one or two chairs in scientific subjects at the universities. But these professors had very few students, for the simple reason that there was no science taught in the schools, which, in turn, was probably due to the fact that there were no professions involving science apart from medicine. But the work of Faraday, Kelvin and Darwin was making a big impression on the minds of thinking people and the importance of the application of scientific knowledge to commerce was beginning to be realized. These lessons were being driven home by the lectures of Huxley and Tyndall and it was felt that some science should be taught in schools both for its value in opening the minds of the boys to new spheres of interest and also to provide a supply of scientists for the service of the state. Accordingly in 1864 a royal commission on the nine public schools recommended that science should form part of the education of all boys. Canon Wilson[1] introduced science at Clifton in 1867 giving as his reasons for doing so that it gave the best training in realizing the meaning of a fact, the meaning of truth. He emphasized that this could only be acquired if the boys actually did experiments themselves and he gave it as his opinion that this sense of a fact could be best acquired by the thorough study of a portion of the subject rather than by a superficial acquaintance with a large part of it. When Prof. F. O. Bower went to Repton as a boy in 1866 he found Dr Pears introducing science along similar lines there, the boys being encouraged to make measurements themselves. But little progress was made in the next twenty years, as classics was strongly entrenched, it led to the majority of the professions in those days, and far more university scholarships were awarded in classics than in any other subject. Consequently very few schools made any serious provision for the teaching of physics either in the way of laboratories or teachers. The subject was openly scoffed at by the masters, it was despised by the boys and disregarded by the parents. The lot of a physics master was a lonely one, cut off from those who were in sympathy with him, with inadequate equipment, and active opposition from colleagues and pupils.

When Sir Cyril Ashford[2] started his career as a teacher of physics in 1888 some progress had been made in that a little science was taught to some boys in all the schools and some of the grammar schools were giving good facilities to boys who wished to specialize in science. But the subject was still the Cinderella of the

curriculum and in a greatly inferior position to that of classics. This obstinate refusal of the classics to recognize any virtue in the newcomer had the unfortunate result that a set of new schools sprung up under the Art and Science Department, whose education was almost exclusively scientific, which was as deplorable as the position in the public schools. Thus an unfortunate hiatus was produced in the national education, the results of which are still apparent in the national life. But another tendency had already manifested itself in the physics which was being taught, in that two schools of thought had arisen. The one, represented by Worthington of Clifton, held that the best training for the boy was to give him a course which was intended for the future professional physicist, irrespective of whether the boy was destined to become a physicist or not. It was a logical course, beginning with the measurement of lengths, areas, volumes, and densities and going on to hydrostatics and so on. Very good results were obtained at Clifton in this course which was based on the syllabus designed by Glazebrook and Shaw for students at the Cavendish Laboratory. Perhaps little ground was covered and the topics were not specially interesting to the boy since they were not specially connected with everyday life, but they were certainly taught something about the meaning of a fact, about the difficulty of accurate measurement, about the way in which facts are classified into laws and about the nature of scientific truth. The work was difficult, it demanded concentration and application, and it did form a good mental discipline. The other school of thought was to be found at Harrow, where there was little specialization in science by the older boys, but the science given to the younger boys consisted of courses of lectures on Royal Institution lines illustrated by demonstrations. The aim of this work, which was possibly suggested by Tyndall, a Governor of the school, was to start from facts familiar to the boy, to show by the demonstrations the methods by which science wins new knowledge, and to show how such knowledge is applied in the various machines with which the boy comes into contact. Already, then, teachers of physics were divided into two schools, those who believed that their main purpose was to give the boy something of the scientific attitude by giving him formal practical work to do and that the information imparted was of little value because most of it would be forgotten when the boy grew up, and those who believed that it was essential to capture the boy's interest by starting from his everyday experience and giving him more information which would be of interest to him. Later on it will be seen that this difference of opinion still persists, but it is probably true to say that most teachers at this early time belonged to the first school, because the schools were largely gateways to the university and so it was natural to frame courses of work which would be suitable for boys going on to the university.

Towards the end of the nineteenth century, there was a feeling that much more opportunity should be provided for secondary education and its natural successor, university education, as there was a large amount of latent ability in the country which was debarred from the benefits of secondary education by lack of sufficient means. So the Education Act of 1902 was passed authorizing the setting up of a large number of new secondary schools, which were to admit a large

number of boys from the elementary schools without charge, the cost of their education being borne partly by the state and partly by the local borough or county council. A considerable number of these schools was established and efficient by the beginning of the Great War and the position of science in them was in no way inferior to that of any other subject. But the courses offered were not in themselves satisfactory. Sometimes a boy studied only physics and no chemistry, or vice versa, or he would study only measurements and heat in physics, ignoring such interesting and important branches as light and electricity. The courses were still of the logical and formal type designed as the foundation on which a university course could be built, although the amount of experimental work done was often too small.

The advent of the Great War finally led the Government to the realization of the vital importance of science in industry and to the realization that all was not yet well with the position of science in the country owing to the lack of the supply of trained scientific men. So another Committee presided over by Sir J. J. Thomson [3] was set up to enquire into the position of natural science in the educational system of Great Britain with special reference to secondary schools and universities. They made two important recommendations as far as schools were concerned. They found that, although science was in every way equal to other subjects in the new secondary schools and many of the older grammar schools, it was still in an inferior position in many of the public schools and also the courses provided in the great majority of the schools were just a foundation on which a university course might be based, although something like 90 per cent of the boys from the day schools, at any rate, had no hope of going to the university. So they recommended that science should form part of the education of all boys at secondary schools and that it should be included in the examinations which had to be passed to qualify for entrance to the universities, to the army, including Sandhurst, and to the professions. They suggested that the course of science should be a complete course based on physics and chemistry, but including some biology, astronomy and geology. This is the first official statement that secondary schools are no longer just gateways to the universities, but that they are intended primarily to give a general education for the future citizen, irrespective of his future profession, and not to give him vocational training for his future profession. They also recommended that the course of general education, which should finish at the age of sixteen, should be followed by two years of specialization for those boys who can profit by it, and that the work should be of the formal logical type designed to lead to a university course. They suggested that boys taking such courses should be allowed to take the university intermediate examination from school or be exempted from it by passing the higher certificate at a suitable standard. The Board of Education followed up these recommendations in 1918 by establishing the various school certificates as the examination designed to test the course of general education, the higher certificates to test the later course of specialization, and they introduced the advanced courses in classics, modern studies and science, in which schools receive a grant in respect of the number of boys taking such courses when they are recognized as efficient. This has been a great help in establishing such advanced courses in science in the

day schools, since it has helped to provide both adequate buildings and a good supply of well-qualified teachers.

To sum up, then, there are three stages in the history of the teaching of science in schools up to the present time. The first stage occupied the last thirty years of the nineteenth century and was concerned with an attempt to get some science taught in the public schools and grammar schools, chiefly to provide university students. The second stage occupied the first fifteen years of this century, when the number of boys receiving secondary education greatly increased and the new secondary schools arose, which immediately placed science on the same footing as other subjects, but still taught science as if all their boys were destined to become professional scientists. The last stage commenced at the end of the Great War and is still in process; it recognizes clearly the distinction between general education to be received by everyone irrespective of their future profession and the more specialized course designed to lead up to a university career. This attempt to teach science to everyone is a gigantic experiment, the magnitude of which has hardly been realized, and it is still largely in an experimental stage, almost the only thing commanding universal assent being that the courses offered should be complete in themselves. The present position of physics teaching as part of a general education will now be considered, starting with the day schools and then going on to the public schools. After that, the present position of advanced courses in schools will be discussed.

§ 3. THE DAY SCHOOLS

The schools have now become conscious of the fact that they have undertaken a new task, that of giving a general education to the future citizen, and that their previous work of laying the foundation of professional courses of study occupies much less time than it did. What is the purpose of this education? What return can the community expect from it? And how is this purpose to be realized?

To start with the first question, why should physics be taught to the future citizen? There are three reasons. Firstly, it opens the boy's mind to new fields of information and to fresh interests; and the seed sown may blossom in later life to an interest in the secrets of nature and to the eternal quest for truth, which will be a source of refreshment amid the burden of daily toil. But while they are still at school, boys take a delight in seeing demonstrated the pressure of the atmosphere; they are fascinated by the sight of a narrow column of water some 5 ft. high supporting five or six boys standing on a board fixed to a motor car inner tube and they take some interest in the explanation; they follow carefully and can reproduce simple demonstrations of the way in which the human eye works, its defects, and the remedies for them; they are thrilled by the beauty and romance of the lines of force around a magnet and will come into school in their own time to make permanent maps of them to keep for themselves; and they can appreciate in some small measure how the discovery of electro-magnetic induction has led to the dynamo and transformer with their distribution of energy in a convenient form to

every part of our country. Secondly the boy can learn from science, and perhaps especially well from physics, a sense of fact and a sense of law[4], two qualities of vital importance to the citizen in modern life with its slogans, half truths and propaganda. It is not claimed that physics will inculcate these things more successfully than classics, for example, but it is contended that with the widening of the stream of secondary education, there are many boys to whom classics makes little appeal and for them science is the better way to acquire this sense of fact and law. Finally education is really just experience; it is the living reaction between the boy and his environment. One way to educate a boy, the way which was the lot of the large majority as recently as a hundred years ago, is to send him straight out into the world and let him acquire his experience there. This may be the ideal education for the very few who can survive "the University of Life", but it is death to the large majority. So a school is provided, which should be a replica of the world, in which the boy obtains such experience as will enable him to survive in the world itself. The part which physics is playing in the world to-day is too well known to members of the Physical Society to need any elaboration here, so it is clear that physics must take its due place in that replica of the world which is called a school. To put it in another way, an educated citizen should know and appreciate a little of the work which other experts are doing. He should have some idea of what sort of work the physicist is doing and what benefits may accrue to the community from it. And he can only do this if he has learned some physics at school.

What time is being allotted to the realization of these aims and what methods are being used? Do they take into consideration the state of the boy's mind and powers referred to at the beginning of this article? It is fair to say that the possibility of teaching the boy a sense of fact and law from scientific work has been in the minds of science masters for some fifteen years now and that some substantial progress has been made in teaching methods in the attempt to realize these ideals. Instead of using the logical method of treatment, in which some law is stated, is verified experimentally, is made the basis of numerical examples and practical applications, the following problem type of treatment is being adopted. Some machine or some striking experimental fact familiar to the boy is quoted and discussed and the way in which the machine works or the explanation of the fact emerges as a problem to be solved. This suggests an experiment, which is then performed either by the master or by the boys themselves, and the law or explanation is derived as a result of the established facts, after some guidance and discussion. There is something of Armstrong's heuristic method in this, but it is not claimed that the boys do research, for it is clearly the master who directs the work. But it *is* claimed that the interest of the boys is aroused by having a problem to solve; they are made to say what the experiment has told them; they are taught to distinguish between experimental error and real variations. And frequently they will find the law themselves, although their expression of it is not formal. This new technique is so important that it deserves to be illustrated by an actual example. If the topic is the moment of a force, the whole subject is approached through various appliances involving moments. The boys are asked to name some levers and the answers will include

such things as scissors, nut-crackers, the brake handle of a motor car and so on. The master can supplement the list with suitable examples of his own, if necessary. The boys are then asked what is the point of such levers and most of them usually see that the idea is to get a big force out of a little one. The problem naturally arises, how does the lever do it? What law determines the extent to which the effort can be magnified? The master then suggests an idealized lever which has all the essential properties of the actual appliances that the boys have mentioned, but with which it is much easier to experiment. This step is a difficult one for boys to follow and is one of the things which can only be acquired by real thought and effort on his part. It is difficult for the boy to see at first that a wooden rod a metre long pivoted at its mid-point, with a weight hanging on either side of the fulcrum, is essentially the same as a pair of scissors. It is the old problem of the boy being so sensitive to the concrete and having so little perception of the abstract. He regards the shape of the scissors as the striking thing; he cannot abstract the important features of the machine from the trivial visual details and so he cannot see the equivalence of the model lever and the pair of scissors. This is why some scientific experiments seem so futile to him; he cannot see that they apply to everyday objects and problems. But it is one of the functions of physics to teach him these things. Again he cannot see that a flywheel rolling down an inclined plane is really like a motor car getting up speed and that it is desirable to experiment with it rather than the motor car because the results are simpler. It is necessary to do a velocity time curve by taking actual measurements with a car first in order to stimulate his interest in the problem, before going on to the flywheel and the inclined plane. To return to the topic of moments, the boys are then sent into the laboratory and provided with simple levers and a selection of weights, they are told to put one weight on each side and to move them about until the lever balances, and to record their measurements. They are then asked to see if they notice anything interesting about the results and, in an average class, there will be a quarter of the boys who will see in some form or other that one force × its distance from the fulcrum is equal to the other force × its distance from the fulcrum. And this is the way in which the law should be quoted at this stage. A number of numerical examples should be set involving the various appliances which the boys have mentioned and in the course of this work it will be noticed that the phrase force × its distance from the fulcrum is always coming in. It is evidently an important term and it can be given a name, the moment of a force, and can be formally defined. It is essential not to give boys a name for an idea until they are familiar with the idea itself. After the quantitative conception is familiar, the boys may be told that the moment really measures the turning effectiveness of the force. This idea should come last, as it is quite the hardest to understand. It should be emphasized that the historical way in which the subject developed is followed here. It is not the so-called historical method of teaching, in the sense that a strict chronological sequence of work by one person or persons is followed, but it is historical in the sense that the ideas are treated in the order in which they occurred historically. Levers were made before the law of the lever was discovered, and from the law of the lever the idea of the moment of a force arose and

then the name was given to it. The name cannot occur before the idea and the idea is usually in men's minds for a number of years before the name is produced. And lastly the physical meaning underlying a numerical quantity emerges. The same sequence can be traced in the second law of thermodynamics, for example. First heat engines were made and used; then the ideal heat engine corresponding to the ideal lever was conceived by Carnot; this led to the discovery of the law governing heat engines, the second law of thermodynamics; its verification by experiment is rather more indirect than in the case of the lever and leads to the emergence of a new idea, entropy, which is first stated mathematically and finally there emerges the physical meaning underlying entropy as the probability of a system or its state of disorganization. When a topic of this kind has been worked out, it is natural to make a logical presentation of the whole subject and this type of presentation is undoubtedly the correct one for the mature mind well familiar with the topic. But a presentation following the historical order of the subject matter is more adapted to the mental state of the beginner and is certainly the right one for boys. Perhaps this is because the development of the boy's mind follows the development of the mental state of the race, just as the development of the embryo follows the evolution of the animal kingdom.

Two further examples may be given of this point, which is of paramount importance since it is concerned with the correct adjustment of the subject matter and method of presentation to the boy's mental state. In studying heat, it is wrong at the beginning to give the boy the idea of heat as a form of energy. To begin with, his ideas of energy are probably very hazy and it is unsound to explain one unknown in terms of another conception almost as little understood. But the boy will accept quite readily the idea of heat as fluid for the simple reason that, like the supporters of the caloric theory two centuries ago, he is familiar with just those facts which suggest a fluid theory. And he will find this notion quite helpful in gaining some understanding why the temperature changes are so different when a drop of hot water is added to a bathful of cold water. It may be stated, in passing, that if the idea of heat as a fluid is suggested as the clue which explains the above fact, the boy's interest is quickened by showing the similarity between the solving of scientific problems and the detective story. This method of treatment has the further advantage that, when the time comes to consider the dynamical theory of heat, the boys can discuss the evidence for that theory and how the caloric theory was replaced by one which could correlate a wider range of phenomena. They will receive valuable training in critical faculty and an insight into the meaning of explanation in science and the status of any scientific theory.

A second example may be quoted from magnetism. When the boys have investigated for themselves in the laboratory the simple properties of a magnet, magnetic induction and how magnets are made and destroyed, it is interesting to ask them what a magnet is or what magnetism is. It is found that many of the boys have already been interested enough to form their own picture or model; they will reply that it is a fluid and that when a piece of iron is magnetized by single touch, some of the fluid goes from the magnet into the iron. This suggestion can be

followed up by asking them what will happen to the strength of the magnetizing magnet in that case and they at once reply that it will get weaker. Indeed many boys ask that question as soon as they see magnetization by single touch, showing that they are thinking about what they are being shown and are forming their own theories. Surely this is education in the best sense of the word, where the boy is using his own mental powers in this way! And he can be shown that his theory is wrong by the well-accepted method that it is contrary to experimental fact. And all this leads on quite naturally to ways of measuring the strength of a magnet. But another answer received to the same question deserves to be placed on record. When this boy was asked what magnetism is, he replied that the magnet had little bits of iron filings in it arranged anyhow before it was magnetized and aligned when it was a magnet. On being asked how this accounted for magnetism, he looked at the master as if he thought he was rather stupid, pointed out that, of course, the iron filings or grains of iron, as he called them the second time, were little magnets and asked if he could go to the board to expound his ideas. He drew an ordinary piece of iron with his grains pointing anyhow (it is interesting that he did not draw closed chains) and then the magnet with the grains aligned. He accounted for the poles occurring as they do in single touch and for their being at the ends. An obvious explanation of this reply is that the boy had seen the Weber theory in a book or had heard some other boys talking about it. But he said that as far as he knew he had never heard of it or come across it before; he had just thought it out. And this boy is very weak in mathematics judged by any university standards, but he has a clear grasp of physical principles. Again another boy had seen a piece of iron put into a solenoid; the current was switched on and the iron was magnetized. He asked if putting a magnet into the solenoid would produce a current in the wire, which is somewhat similar to the question which Faraday asked himself and which led him to his discovery of electro-magnetic induction. These spontaneous questions and ideas do lend support to the principle that subject matter should be unfolded to the boys in the same way as it developed historically. The facts with which they are familiar are the same as the only facts known to the earlier investigators, and the theory historically put forward to account for them is probably the one which is most in harmony with their mental state. But it would be unwise to carry this principle to the extreme limit and to include no modern concepts. In teaching electricity, it is important that the boys should be given some idea of the electron. After all, the electron theory will explain all the simple facts of electrostatics and beyond that the boys must be told that there is further evidence which cannot be given in detail. Such allusions to modern work and concepts link up the classroom with the stream of living knowledge which is constantly flowing from the universities and research laboratories and bring life and freshness to the classroom itself. It is claimed that the sense of fact and law is inculcated by this approach to the subject matter to be studied. Its ideal of quoting no fact unless the boy either establishes it for himself in the laboratory or sees it settled by experiment in the lecture room, its attempt to present the work as a set of problems connected with some piece of everyday experience or some striking fact which has been demonstrated in order

to maintain a high level of interest, its effort to bring something of the research attitude into the classroom both by the spirit in which the work is tackled and by its allusion to modern work, all contribute to this end and it may be hoped that something of this sense of fact and law may be transferred to his everyday experiences as he grows up.

The other two aims of teaching physics to the future citizen are to open fresh fields of interest to his mind and to give him some idea of the function of the physicist in society. These ideals can best be realized by a suitable selection of subject matter and it is felt that the formal course, which aims at being the foundation on which a university course can be built, is compelled to devote so much time to acquiring a technique, that there is little or no time left to acquiring any interesting information with that technique. An attempt is accordingly being made to strike a more reasonable balance between technique and information and it will be seen that the old question of the rival claims of mind training and information is still unsolved. A concrete example will illustrate the way in which the problem is being tackled. In the more formal course, a good deal of time is devoted to measurements in magnetism and an attempt is made to deal with them logically in the sense that no quantities are used whose measurement has not been accomplished by the boys or explained by the master. For example, the value of H is not assumed, because its measurement is too difficult at this stage. So the work is confined to the comparison of fields and moments, both broadside and end-on positions being studied. This leaves very little, if any, time for the discussion of results of interest which have been obtained recently. A more even balance is maintained between technique and information if it is assumed that $H = 0{\cdot}18$ gauss, although no idea of the way in which it is measured can be given, and actual fields and moments are measured, the broadside-on position being ignored. There is then time to give a simple statement of the quantitative definition of susceptibility and coercive force and to discuss the striking new magnetic alloys such as permalloy, perminvar and the cobalt steels, which have been made recently. Also a lesson can be devoted to the origin of the magnetism of the elementary magnets, which will have been mentioned in Weber's theory of magnetism. The boys ask this question spontaneously and take pleasure in linking this problem with the magnetism of a current in a coil, which will have been demonstrated to them. And there is no reason why they should not be told that elementary magnets owe their magnetism to electron spin, and the Richardson gyro-magnetic effect may be quoted in support of this view that magnetism is associated with rotation, the conservation of angular momentum being demonstrated with a bicycle wheel and a turntable. And they can be told that the elementary magnets are domains, giving them some rough idea of their size. They may not understand all about these things, but which of us understands anything completely the first time? The important thing is to give the boys a glimpse inside the inner chambers of science, to let them feel the fresh breeze of new knowledge around them. This treatment will show the boy how the physicist plays his part in the life of the community and it gives him a hint of the eternal search for truth which is being everlastingly pursued and which may be

a source of refreshment in after years when he is bearing the heat and burden of the day.

This attempt to strike a new balance between technique and information and to select the subject matter so as to give the boy both some conception of the part the physicist plays in life and of the eternal search for truth is being realized to some extent by the introduction of general science. The syllabus suggested in an interim report[5] of a sub-committee of the Science Masters' Association cannot be quoted here, but two characteristics may be noticed. It draws its subject matter from a wide field of topics and grades the treatment in accordance with the mental state of the boy. The work of the first year is largely qualitative, any quantitative work involving only simple arithmetic. It rightly includes a course of sound and light, both of which are of real interest to boys and give valuable training in scientific method. Measurement is first introduced in the attempt to find a unit of heat; specific heats are treated very simply, as it is not necessary to measure the specific heat of copper or, say, a common mineral to three significant figures in order to discuss why the water in a swimming bath takes so long to warm up after it has been newly filled, or to account for the difference between a continental and an oceanic climate. And the chief reason for going into the calorie in a course of this kind is to be able to measure the calorific value of a fuel and of foods. The work in mechanics is designed to lead up to the principle of work and the conservation of energy, two principles which should be known to every citizen in these days of mechanical civilization. Some reference is also made to Newton's laws of motion and their explanation of the motion of the planets in a very simple way. No electrostatics is introduced, the idea of potential being introduced by analogy rather than logically. This omission of electrostatics does not command universal assent by any means, for the subject matter interests boys and potential difference is so fundamental in the electric current, that it is a pity to leave the boy merely with an analogy. But some of them cannot understand a logical treatment at this stage. The bulk of the time in electricity is devoted to explaining the mode of action of the common electrical instruments, such as ammeters, voltmeters, motors, dynamos, telephones and transformers and to discussing the laws relating the transformation of electrical energy into other forms of energy.

This scheme of work represents the best attempt which has so far been made to provide a course of physics suitable for the future physicist. Previous schemes of general science were rightly criticized because they lacked mental discipline and merely gave the boy a superficial acquaintance with much of the subject, largely second-hand, and with no idea of how the scientist gains his knowledge. There is more discipline in this course and more first-hand information is obtained by the boys, but general science has not gained anything like universal assent in the schools yet. All secondary schools provide a four-year course in science between the ages of twelve and sixteen and many schools are now adopting general science in the first two or three years of that course, taking a more formal course for the last year or two years. There is a feeling that the course still lacks sufficient mental discipline and that it is not a sound foundation on which to base the advanced

course which is to follow in the case of some boys. While the future citizen rightly bulks largely in the aims at the present time, the interests of those boys who wish to go to the university must not be altogether forgotten.

§4. THE PUBLIC SCHOOLS

The public schools may be defined as a small set of schools of non-local character, usually boarding schools, which do not receive any financial assistance from the state and so are free to develop along their own lines. The system of education offered in these schools, although it is secondary education, is somewhat different from that provided in the day schools, in that it commences with some five years at a private preparatory school until the age of 13+, when the boy goes to a public school for three to five years. His course at the preparatory school is English subjects, Latin, French, and mathematics, little or no science being taught. The average boy will spend a further three years over his general education at his public school and will then specialise for one or two years. But there is much more range of ability entering the public schools than the day schools, and some boys will take only two years to complete their general education, exceptional boys only one year, although there are strong arguments for regarding this as educationally unsound. Again other boys take four years to complete their course of general education and some never complete it at all in the sense that they never take the school certificate.

The same divergence of opinion is to be found among the public schools as among the day schools concerning the best course of physics for the future citizen, although there is agreement as to the aims to be realized. Some prefer the more liberal course provided by general science, but there is quite a strong feeling that the standard required in the examinations in this subject is so low that it lacks mental discipline. So a number of schools prefer the rather more formal course leading to physics with chemistry as a joint subject in the school certificate, contending that the syllabus required by the examination permits of quite a liberal treatment of the subject matter. This preference for the more traditional course in some schools is supported by the fact that a greater proportion of their boys are destined for the university than in the day schools. In some day schools as few as 2 per cent of the boys go to the university, while something like 50 per cent of the boys at a public school go to the university, Woolwich, Sandhurst, or agricultural colleges. Some schools are trying the experiment of general science for the boys who will specialize in arts, preferring physics with chemistry for the boys who are to specialize in science. Some administrative difficulties arise from the varying abilities of the boys at entrance. Some schools provide a three-year course in physics to be taken by those who will take three years to complete their general education, making no provision for the abler boys who will take only one or two years to pass the school certificate. Others provide a two-year course leading to the school certificate, so that the abler boys may take science right from their entry into the school, and give the weaker boys courses in the more elementary parts of light and sound in their first year.

There are three unsatisfactory features about the position of physics in public schools. The total number of boys in these schools may be small, but as many of them are destined to occupy important positions in public and commercial life, it is vital that they should be alive to the importance of physics to the community. Sir J. J. Thomson's Committee reported in 1918 that it was unsatisfactory that science was not taught to the boys on the classical side at the public schools and recommended that these boys should all be taught some science. Most public schools still provide a separate classical side even during the course of general education, which differs from the modern side in that the boys take three foreign languages, French, Latin and Greek, instead of two, French and Latin or German, as on the modern side. They do this at the expense of science. The large majority of the able boys go on to the classical side on entering the school, because the majority of entrance scholarships into the school are awarded for excellence in classics. The position at the public schools has improved somewhat since the recommendation of Thomson's Committee in that some schools do insist on boys on the classical side taking a course in science, although they may not be examined in it. But there are still a number of schools in which it is possible for boys reading classics to do no science at all during the whole of their school career. This is wrong for two reasons: firstly, they are only half educated and are unfit to occupy positions of responsibility in the state, and, secondly, some of these boys may have an interest in, and aptitude for, science and might have been able to render their best service to the state in some form of scientific work. This contention is borne out by the fact that, at schools where science is taught to boys on the classical side, some of those transfer to the science side when they wish to specialize and sometimes gain university scholarships.

The second unsatisfactory feature is that the course of general education in physics is too short in the case of the average boy and that it is started too late. Consequently the subject matter is somewhat out of step with the boy's mental development; parts of the subject designed for boys at an age when they are primarily interested in facts have to be presented when the boy has reached a later age and is more interested in practical application and machinery. The trouble arises from the fact that the boy receives his education at two different schools and he is not yet taught any science at the preparatory school. Sir J. J. Thomson's Committee recommended in 1918 that some nature study or physiography and the simple measurements of lengths, areas, volumes and densities with the use of the metric system should be taught in preparatory schools; that science should be a compulsory subject in the common entrance examination which is taken by all candidates who wish to enter a public school; and that science should be a subject in entrance scholarship examinations to the public schools carrying as much weight as other more traditional subjects. Although twenty years have passed since these recommendations were made, practically nothing has been done to introduce them! Science is not yet a subject in the common entrance examination, little science is taught in the preparatory schools, and the subject carries little weight in the scholarship examinations.

The third unsatisfactory feature of physics in public schools is that the average boy is not well enough equipped in arithmetic to be able to tackle the simpler quantitative parts of the subject with confidence. Instead of arithmetic being a tool which can be used in physics, instead of the boy being able to handle multiplication and division in decimals with facility and accuracy, it is a thing over which he must spend mental energy and so he has an insufficient amount left to devote to understanding the physical principles which are being presented to him. It is not the case that these boys are lacking in intelligence; they can solve problems of a qualitative kind, or of a quantitative kind if the numbers are very simple, but they fail if they are given the sort of numbers which might occur in an actual practical case. The very same complaint appears in the report of Sir J. J. Thomson's Committee, but the situation seems to be very much the same to-day as it was twenty years ago. This weakness in arithmetic and the absence of any serious teaching of science both arise from the small size of the average preparatory school, which makes it impossible to have more than one really well-qualified master on the staff. He must be a classic, since excellence in classics is essential to gaining entrance scholarships into the public schools, and the gaining of such scholarships is one of the hallmarks of success in a preparatory school, the others being good games and the right social tone. It is the exception to find a well-qualified mathematics teacher at a preparatory school. Where there is one the standard of mathematics is good, but in most cases it is deplorably low.

It is essential that these unsatisfactory features of physics in the public schools should be remedied and the time is ripe for their solution, as it is clear that science is going to be more, rather than less, important in the solution of the grave problems which lie before the state and, in turn, the public schools and preparatory schools may be more willing to adopt new ideas, as they are already beginning to suffer from the decrease in the number of boys available for secondary education on account of the falling birth rate and also from a decrease in their own numbers on account of the decrease in wealth of the upper classes. The three problems are really all bound up with one another; science is excluded from the preparatory schools and mathematics teaching is weak, because it is essential to have good classics teaching to get the best boys scholarships into the public schools. And boys on the classical sides of such schools do no science and spend sometimes as little as one year on a course of general education, because they must get an entrance scholarship to the university. But the day schools have shown that it is quite possible to get university scholarships in classics without starting Greek at the age of eleven and while giving the boy some science during his course of general education, and the education of public school boys would probably benefit if they learned less languages while they were at a preparatory school and did some serious science instead. Is it educationally sound and in harmony with the boy's mental state that he should be learning three foreign languages, two of them dead languages, at the age of twelve or earlier? It is well established that he has a keen interest in facts about nature at that age and he is starved of this at the preparatory school and then fed with it at a public school when his appetite for it has gone. The remedy is clearly in the hands of the

headmasters of the public schools themselves and it is to be hoped that they may be persuaded to act. Firstly, the practice of pushing boys through the school certificate at the age of 14+ after only one year of general education, which contains no science, should cease; it still happens at a number of public schools. We do not ask for the abolition of separate classical and modern sides as was recommended by Sir J. J. Thomson's Committee twenty years ago, but only that boys on the classical side should do a serious course of science, even if they are not examined in it. Secondly, the preparatory schools should be informed that science will be a subject in the common entrance examination in 1940 and that it will be compulsory in 1943. This would allow them ample time to make arrangements for the teaching of the subject. Finally, the examination of mathematics in the common entrance examination should be designed so as to cover a smaller number of topics but to demand more accuracy and facility in arithmetic.

§ 5. ADVANCED WORK IN SCHOOLS

What is called advanced work in schools covers the same ground as the first year course in a university and is intended to prepare the boy for a university intermediate examination or something equivalent to it. The members of the Physical Society are quite familiar with the contents of such a course, which is designed to give the student an elementary knowledge of the principles of physics such as are needed either for a more specialized study of physics itself or of other sciences such as chemistry or physiology. The aim now is quite different from what it has been at the stage of general education; it is now to train future scientists and perhaps even future physicists. The course is rightly modelled on academic lines, the subject matter being developed logically.

Such courses have been given in some of the grammar schools and public schools for the last forty years chiefly as a preparation for university scholarships and external intermediate examinations. But they were hampered both in masters qualified to teach the work and in laboratories and lecture rooms in which it could be done. A great step forward was taken in 1918 when the Board of Education introduced a system of grants for schools which provided advanced courses in classics, modern studies, or science and mathematics. The financial assistance was conditional on the work reaching a sufficiently high standard and sufficient number of boys taking the course, and in the case of science it has been a great help in that it has provided better laboratory equipment and more teachers for these courses. The laboratories in many schools are now quite as well equipped as the elementary university laboratories, and even better than in those universities where all available money is spent on research! The classes are considerably smaller than those in the universities and some of the masters teaching advanced work in schools are quite as well qualified as university lecturers, having been attracted to schools by the fact that the prizes in university work are awarded almost exclusively for excellence in original work, little reward being given to excellence in teaching. The quality of the work done in these courses is good, as is shown by the keen competition for

entrance scholarships in science to Oxford and Cambridge, which are considerably more numerous than they were twenty years ago; many boys pass the intermediate science from school each year, and many Cambridge colleges are insisting on medical students passing the physics, chemistry and biology of the first medical examination before they go up to Cambridge, thus implying their confidence in the schools to teach this work well. It may be added that a number of schools succeed in getting their boys through the organic chemistry of that examination as well. Sir J. J. Thomson's Committee recommended that boys should be allowed to take intermediate and first medical examinations from school and it is gratifying to see that this recommendation has become an accomplished fact.

As the schools are laying the foundations on which the universities have to build later, it is essential that there should be some contact between the two bodies and that they should understand one another's problems and co-operate to produce the best training for their pupils. During the last two years at school, the boy must be introduced to a logical treatment of physics and trained to think logically. Since the primary aim of physics was to reduce all phenomena to mechanics, the course must include a sound training in the principles of statics and dynamics. There is still disagreement as to whether it should be taught by the physicist or mathematician, but one solution is that the subject should be in the hands of the physicist while the fundamental principles such as the principle of moments and the laws of motion, which need experimental support, are being established, while the mathematician should take over when the subject develops into the mathematical application of those principles. In practical terms, the subject might be taught by the physicist in the first year of specialization and the mathematician in the second year.

The schoolmaster must realize that the university expects to receive pupils who can make themselves work and can learn to a considerable extent from lectures and books. So he must train his pupils in self-discipline, teaching them to keep themselves up to the mark; punishment must disappear from the landscape to be replaced by friendly but firm advice. Teaching, with its questions and discussions, its monthly tests as a check on the boy's progress, must be gradually replaced by lecturing, in which the boys will be trained to take notes, and terminal examinations. The boys must also be trained to learn the easier parts of a subject from the book by themselves, some general advice being given as to how to learn from books. This can be introduced by planning the two-year course so that the various branches of physics, heat, light, sound, magnetism and electricity are each treated in the first year up to a certain standard and are then repeated in the second year to a higher standard. If a separate text-book is used in the second year, the boy can be encouraged to learn from it some of the topics which he has done in the first year and which he will not yet know thoroughly.

The university lecturer must realize on his part that the atmosphere of a school is fundamentally different from that of a university and that boys at school have not the same stability of character as undergraduates. So even at the advanced course stage, it is necessary to make the work interesting. It is not enough to take the

attitude that the boy has chosen to do physics; this is physics, if he does not like it, then he should specialize in some other subject. Now the advanced work at schools is, after all, only the elementary principles of the subject; they are well known and well established. It is not easy to arouse interest where there is no uncertainty, no new work. Part of the appeal and interest of modern physics is its uncertainty, its freshness. Something of this freshness and vitality must be brought into this elementary work if it is to interest the boy and provide a sound training on which to base the more advanced work. If the subject is presented in logical sequence, it is difficult to arouse interest and the presentation is apt to become dull and even dogmatic. It fails to train the boy's critical faculty. It is best to illustrate this point by a concrete example. If the subject is light, geometrical optics is apt to be dull if it is presented as a set of mathematical results. But it can be brought to life if the boys are told why these results were ever investigated. Do not just give the experiments themselves, but discuss why they were done, what problems inspired them. Try to show the subject as a living thing, growing as the result of practical demands and theoretical problems. Geometrical optics arose in the attempt to analyse the behaviour of lenses and mirrors, to probe and improve the working of optical instruments. And let problems of this sort be set: a projection lens is required to produce an image 8 ft. × 8 ft. of a standard lantern slide 3¼ in. × 3¼ in. on a screen 12 ft. from the slide. Calculate the optician's specifications of the lens if it is to be made of glass of refractive index 1·50. This is a practical problem which will interest the boy and show him the point of geometrical optics. It can be applied to chromatic aberration and the other lens aberrations, illustrating how the number of variables finally diminishes as the attempts to produce a more perfect image proceed. When the nature of light is discussed, it raises interest and gives most valuable training in critical faculty if the corpuscular theory is discussed as well as the wave theory. The method of approach is really this and it can be applied to all branches of physics. When the boys have learned just those facts which were historically known at the time that serious theories were put forward, let the theories be put to the boys themselves and let them discuss them. The work at once comes to life; there is a problem; there is uncertainty; there is a reason for doing experiments. If only one theory is expounded to the boys, there are two disadvantages. Firstly, they can learn it off by heart, knowing that they are learning the right thing, the thing which will get them through the examination, which is the way to spiritual destruction, and the master can compel dogmatic acceptance of the theory. Secondly, they miss the stimulus and discipline of really having to think for themselves, whether an explanation is right or not; they just rely on the authority of the teacher, instead of on the experimental evidence and the force of pure reason. To return to the concrete case, the honest keen boy never quite takes to the explanation of reflection on Huygens's principle, or refraction for that fact. It seems to him to be a rather involved way of explaining a simple fact and the honest boy says so. He is much more attracted to the idea of an elastic impact of a corpuscle, which leads simply to the laws of reflection. When the facts of light have been presented to him, he will see that the corpuscular theory has natural explanations to offer of

rectilinear propagation, reflection, refraction and colour, but it fails over simultaneous reflection and refraction and the velocity of light being independent of the temperature of the source. The wave theory, on the other hand, has no explanation to offer of straight line propagation and it makes a gigantic assumption of the ether to explain why light travels through empty space. Neither theory, then, is completely satisfactory. What is to be done about it? Here the boy is confronted with the situation which faced the original investigators of the seventeenth and eighteenth centuries and with the type of situation which may face him in a much smaller range of facts if he does research in the future. He learns, then, that an important function of a theory is to suggest new facts and that the wave theory is more successful in this line than the corpuscular theory. So a discussion of the properties and theory of waves follows next and suggests a search for interference and diffraction in light. It is not a case of doing interference next because it comes in the book; it is done to settle that conflict between the two theories. And is it not right to insert a brief account of Newton's explanation of Newton's rings? Is it not true to say that this is an admirable example of the true scientific spirit of compromise, of being guided solely by the facts? Newton felt that rectilinear propagation ruled out the possibility of light being waves, but after he discovered the rings, he saw with uncanny intuition that the periodic recurrence of the rings was an indication of some periodicity in the light itself. So he introduced waves into his theory of light. It sounds contradictory, but was not Bohr's original theory of the spectrum of the hydrogen atom full of glaring contradictions? But who would dare to dismiss that with the contemptuous references with which Newton's explanation of Newton's rings is dismissed in some text-books? The boy will not find this work an undue tax on his mental powers for they will be stimulated by the problems raised and some of the explanations on the corpuscular theory are so simple and beautiful that they are easily understood and remembered. The boy will finally learn that weight of evidence favours the wave theory, which triumphs not because of any crucial experiment (has the quantum theory been vindicated by any one experiment and could it be disproved by one experiment?) but because it can correlate a wider range of facts than the corpuscular theory. He will learn the valuable lesson that any scientific theory may be revised in the light of wider evidence, but that this is not a question of knocking down what has been built up; it is more in the nature of an extension, for the new theory must reduce to the old one in the sense that it leads to the same results, when applied to the original restricted set of facts. The wave theory must lead to the same laws of reflection and straight line propagation as the corpuscular theory when the wave-length of the light is small compared to the apertures or obstacles in use.

It has already been mentioned that the physics given to the future citizen should be vitalized by linking it now and then with the work continually issuing from centres of research. This is also true of the specialized courses in physics. The university lecturer may say that this is out of place in a logical course; the school should lay the foundations soundly, and the university will add the modern work. But the pupil is not a machine; he is a living being, keen on his subject; he will have heard

of this work, especially if he is encouraged to glance at *Nature*, *Discovery*, or to read such books as Andrade's *Mechanism of Nature* or the *New Chemistry*. He wants to know something of this work and it is right that his curiosity should be satisfied to the extent that such parts of it may be introduced as he can properly understand. This contention may be supported by returning to the case of light once more. After polarization and double refraction have been dealt with, the simple facts of the photoelectric effect may be introduced; a metal plate acquires a positive potential of about a volt when exposed to ultra-violet light; this is due to the emission of electrons. Then the question may be put to the class for discussion: how will the number and speed of the emitted electrons depend on the intensity and frequency of the incident light on a wave theory? After the ideas of the class have been put into a systematic form (they *do* put forward ideas quite similar to those suggested historically), further facts may be brought forward, such as the threshold effect and the independence of the velocity of the emitted electrons on the intensity of the incident light. There is a serious failure of the wave theory to account for the facts in a natural way here. The attempts to find a way out by postulating that photo-electric emission is a kind of resonance effect and the trigger hypothesis may be discussed. Einstein's quantum theory of the photoelectric effect should then be introduced, when the pupils will be fascinated by the reappearance of the corpuscular theory in a new form. Once the fundamental physical principles of the atomic nature of radiation are understood, the equation relating the frequency of the incident radiation and the velocity of the emitted electrons is easily derived, the mathematics being so simple as to be no obstacle. A very brief account of the experimental verification of this prediction by Millikan should be mentioned. Work of this kind is an admirable training of the boy's critical faculty and it is not too hard for him. It certainly differentiates between those who learn by heart or are content to accept things on the authority of the teacher and those who are genuinely interested in their work and will accept only established fact and the light of pure reason. It does give the pupil some sort of training in the solving of problems and if he cannot do that, he is no use to anyone. The work may be concluded by a brief qualitative treatment of the Compton effect, in which the underlying ideas are so simple. If radiation is corpuscular, then the scattering of radiation by free electrons may be regarded as an elastic impact between the quanta and electrons. Since the electrons receive some energy in the process, the quanta are scattered with diminished energy, that is, decreased frequency. The mathematical equation for the increase in wave-length can just be quoted and the verification can be discussed briefly. The work concludes with a question: how can this dual nature of radiation be explained? It can be pointed out that some answer must be obtained to this problem, as it strikes to the root of things. Perhaps this point may be driven home by mentioning that there is evidence that electrons, hitherto regarded as particles, sometimes behave as waves. Any course should end in this way; it is essential to inculcate in the pupil's mind that science is a set of successive approximations to the truth; that, as new ground is won, fresh problems open themselves for solution, new facts are discovered demanding ever wider generalizations. This

will educate and train the student for the changing world in which he finds himself to-day and which is surely a reflection of the major contribution of science to present day civilization.

How is the university lecturer to build on this course? The pupil comes to him with his logical faculty more developed, with his character more disciplined, his interests more fixed. He is now ready for the strictly logical treatment and will delight in it. To take a concrete example, a course of light should begin with the equation of wave-motion, and all the properties of waves should be deduced logically from it. First the properties of the waves when impinging on instruments such as mirrors and lenses large compared to the wave-length should be considered and the familiar lens and mirror equations and properties appear from a new angle. Then the phenomena of interference and diffraction follow in their logical sequence, the results being deduced mathematically and the experimental verifications being discussed as they arise. Naturally the topics will be carried further than at the first treatment in a course at school. Interference, for example, will not only be discussed for its bearing on the nature of light, but it will be shown how light waves become a tool with which small distances can be measured and finally have become a standard of length. This illustrates how scientific research just pursued for its own sake does furnish results of practical importance. The university and school courses will cover the same ground to some extent; this is desirable and necessary for the development of the pupils' minds, but there need be no loss of interest if the method of approach is different as is here suggested. The same subject matter is treated but in quite a different way and the freshness is maintained.

The assertion is made now and again that the examinations set to test the two-year course of specialization at schools are too hard; the questions are said to demand too advanced and too detailed a knowledge of the subject. Taking a broad view, this cannot be maintained in the case of physics. For the last twenty years the public schools have been gaining scholarships at Oxford and Cambridge in science with the following type of training: two years of general education, in which about a fifth of the time is devoted to science, followed by two years and one term specialization, in which about three-quarters of the time is given to science. Considering that these boys did no science before they were thirteen, it cannot be said they have been crammed, and they are quite capable of assimilating the work necessary to reach a high standard in the examination. It may be true that some weaker candidates are crammed in order to force them just up to exhibition standard. Naturally they fail to fulfil their early promise when they go to the university, but does not the responsibility for this lie with the master? He can judge a boy's ability fairly well when the boy begins to specialize, and if he is not up to the standard the master should refuse to cram him and have the courage to give his reasons to the parent. It is time that masters realized that they have a service to the state before their own particular school and while it may be good for the school to have a long list of scholarship winners, it is not in the best interests of the state to have boys crammed to win scholarships, of which they cannot take advantage, nor make any suitable return to the state for their university education. When schoolmasters

begin to take a national view of their work and when parents and local authorities realize that the number of scholarships a school wins is largely due to the number of able boys which enter the school rather than to the excellence of the teaching in the school (for what teacher will claim that he can turn the average boy into a university scholar?), then the problem will solve itself. Again it is true that questions may sometimes be set on modern topics, such as television, X-rays or transmutation. But this does not mean that these topics are to be included in a school course. It is rather to see if the boy has enough interest in physics to read for himself, and looked at from this point of view the questions are quite suitable.

§ 6. CONCLUSION

The character of secondary education has undergone a profound change in the last forty years, in that it formerly trained pupils for the university whereas it now trains them for the duties of ordinary citizenship. Like most things in this country, the change has come about gradually and it is only in the last few years that the schools themselves have become thoroughly conscious of the gigantic task which they have undertaken. There is agreement that physics should form a substantial part of the education of the future citizen and as to the reasons why it should do so, but there is not yet agreement as to the type of course which will best realize those aims. During the course of a discussion on "Education for a Changing World" at the Cambridge meeting of the British Association, one member remarked that he had attended meetings of the Association for fifty years and none of the suggestions he had heard this year were new; they were all being made fifty years ago. This statement expresses a real need in education; it is time that a serious attempt should be made to bring some of these suggestions into force and to test them scientifically. Valuable research is being done in education in the development of intelligence tests, for example, but much more work of this kind is needed, if education is to make any real progress and is to justify the large expenditure of money on it. To take one concrete case in the present discussion, it has been stated that one of the aims of teaching physics to the future citizen is to give him a sense of fact and law. It is essential to find some way of measuring these senses and to undertake large scale research, with large enough groups of persons to make the results statistically significant, to see if these senses are being enhanced by the work done in schools and to vary the curriculum and methods of teaching until the best results are obtained. Another aim is stated to be that physics gives the future citizen fresh interests and opens his mind to the beauty of natural law and should give him an interest which will be a source of refreshment to him amid the toil of everyday life. This point could be tested by large scale investigations; it should not be difficult to find out how many secondary school pupils do maintain any interest in physics in after life. It is only if scientific research of this kind is undertaken that any serious progress in education for the citizen will be made and that the present unsatisfactory position of the same suggestions being made for fifty years will be abolished.

Although recent investigations have shown that examinations are not a reliable test of a student's knowledge of a subject, they do determine to some extent the type of work done and the number of pupils taking a particular course. General science suffers at the moment from the examinations in it being distinctly less satisfactory than those in the more traditional courses. In some cases the papers are easier than those in physics and chemistry and so have led to general science being regarded as a soft option to be avoided by all self-respecting pupils who wish to face difficult work. In other cases the questions are unsuitable and are not well worded. This is not a reflection on the examiners, but it is fair to say that no one should be expected to examine in this subject who has not taught it or has not some first hand acquaintance with the teaching of boys. It is difficult to realize how little discrimination boys possess. For example, a boy was recently asked in a practical examination to measure the resistance of a wire with a Wheatstone bridge. He was provided with the usual metre bridge, but as this pattern had always been called a metre bridge, and although it had been quoted and explained to him as a particular case of the Wheatstone bridge, he asked if he was allowed to use the metre bridge. Once again it is a matter of distinguishing between essential and detail; he could not see that there is no essential difference between resistances in the form of a straight wire and resistance bobbins. So the questions need to be set with attention to this point; they must be definite questions and the scope of the answers required must be made clear. Some examination boards do appoint schoolmasters as examiners or do ask schoolmasters to look over the questions set by university examiners to see if they are suitable for boys. But something more radical is required than this. General science papers will not be satisfactory until they are set by schoolmasters. It may be objected that it is out of the question to have the schools examining their own work; some outside body must do this. But the universities examine their own work; there may be an internal and external examiner, but the external examiner is still a university lecturer. In fact, the universities settle what they should teach and they examine it themselves. They have earned the right to do so, because, by many years work, they have convinced the community that they have a real contribution to make to the welfare of the state and that only those candidates with a high standard of ability will be granted a pass in their examinations. The schools will never play a real part in the life of the state until they win the same position for themselves. They alone teach the future citizen; they alone must decide what is the right course for him; and they must set a suitably high standard for a pass in their examinations, so that those who succeed can and do render efficient service to the state.

Reference has already been made to the fact that recommendations of Sir J. J. Thomson's Committee made twenty years ago to improve the position of science in the public schools have not yet been acted upon. It is to be hoped that science may be made a compulsory subject in the common entrance examination so that the teaching of science in the public schools may be more in tune with the boy's mental development. If the preparatory schools find it impossible to undertake any serious science, then an alternative solution might be to lower the age of entry

to the public schools to 12, when a systematic science course could be planned. There is no logical justification for the present break at the age of 13+ in the boy's education.

Finally, the quantity and quality of advanced work done in schools has increased since the Board of Education established grants for advanced courses in 1918 and the work which used to form the first year of a university course is being done more and more in the schools. This was in accordance with a recommendation of Sir J. J. Thomson's Committee and it is to be hoped that this tendency may continue. This will only be so provided that the schools can reach a standard and give their pupils the type of training which the universities require and since the two bodies are to some extent on common ground here, it is desirable that there should be more contact between them. Much has already been done by the willingness of the universities to provide hospitality to the Science Masters' Association, by the kindness of professors and lecturers in delivering lectures at their meetings and contributing articles to the *School Science Review*, and by the Physical Society in arranging discussions. It is to be hoped that more contact of this kind may be established; would it not be possible to arrange for a discussion of the Physical Society and Science Masters' Association on the teaching of science in sixth forms at schools and intermediate classes at universities? It is to be hoped that this would lead to an increased respect between university lecturers and schoolmasters, a respect for one another's aims and a sympathy for one another's problems, an increase in the quality of advanced work done in schools, and a more general recognition of its value as exempting from the first year's work in the university. For the advanced work in schools is of great importance to the life of the whole school, forming a link with the universities from which the stream of living knowledge is continually issuing. It is this stream of new knowledge which has kept the great public schools and day schools alive for the last five centuries and which must play its part in sustaining the much wider secondary education which is being attempted at the present moment.

REFERENCES

(1) Westaway, F. W. *Science Teaching* (1929).
(2) Presidential Address to the Science Masters' Association (January 1938).
(3) *Natural Science in Education* (H.M.S.O.).
(4) Norwood, Sir Cyril. Presidential Address to the Science Masters' Association. *School Sci. Rev.* **13**, 169.
(5) *The Teaching of General Science* (Murray).